ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Μέθοδοι αυτόματου εντοπισμού ορισμών σε συλλογές εγγράφων*»

Γεράσιμος Λάμπουρας

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2008

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Στη διάρκεια προηγούμενων εργασιών, κατασκευάστηκε ένα σύστημα το οποίο δέχεται ως είσοδο έναν όρο και εξάγει ορισμούς του όρου από ιστοσελίδες που επιστρέφει μια μηχανή αναζήτησης του Παγκόσμιου Ιστού. Πιο συγκεκριμένα, το σύστημα δημιουργεί αυτόματα, χωρίς ανθρώπινη επίβλεψη, παραδείγματα εκπαίδευσης μιας Μηχανής Διανυσμάτων Υποστήριξης, την οποία χρησιμοποιεί κατόπιν για να κατατάσσει τμήματα των ιστοσελίδων ως ορισμούς ή μη-ορισμούς. Στη παρούσα εργασία, εξετάζουμε τρεις νέες μεθόδους αυτόματης παραγωγής παραδειγμάτων εκπαίδευσης της Μ.Δ.Υ. Δοκιμάζουμε, επίσης, τεχνικές αφαίρεσης πλεοναζουσών υποψηφίων απαντήσεων του συστήματος. Ακόμα, μελετάμε τα κορυφαία σχετικά συστήματα που έχουν παρουσιαστεί διεθνώς και συγκρίνουμε τα αποτελέσματά τους με αυτά του δικού μας.

1. ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Τα συστήματα ερωταποκρίσεων (question answering systems) φιλοδοξούν να αποτελέσουν την επόμενη γενιά στις μηχανές αναζήτησης εγγράφων. Σε αντίθεση με τις υπάρχουσες μηχανές αναζήτησης, λαμβάνουν ερωτήσεις φυσικής γλώσσας και επιχειρούν να επιστρέψουν σύντομες πιθανές απαντήσεις, αντί για καταλόγους σχετικών εγγράφων. Αν και τα συστήματα ερωταποκρίσεων (π.χ. για βάσεις δεδομένων) είναι μια από τις αρχαιότερες ερευνητικές περιοχές της Επεξεργασίας Φυσικής Γλώσσας [AnRi95], ιδιαίτερο ενδιαφέρον για τα συστήματα ερωταποκρίσεων συλλογών εγγράφων (π.χ. για αρχεία εφημερίδων ή τον Παγκόσμιο Ιστό) εμφανίστηκε στα τέλη της δεκαετίας του '90. Ήδη έχουν ανακοινωθεί πολλά ελπιδοφόρα αποτελέσματα στον τομέα αυτό, πολλά στα πλαίσια του Question Answering Track του TREC (Text Retrieval Conference) και άλλων συνεδρίων.[1]

Στη διάρκεια προηγούμενων εργασιών των Μηλιαράκη [Mi03], Γαλάνη [Ga04], Γιακουμή [Gi05] και Λάμπουρα [La06] κατασκευάστηκε ένα σύστημα το οποίο εξάγει τις απαντήσεις του από τις ιστοσελίδες που επιστρέφει μια μηχανή αναζήτησης του διαδικτύου. Το σύστημα επικεντρώνεται στις ερωτήσεις ορισμού (π.χ. «Τι είναι ο μυστικισμός;»). Συγκεκριμένα, δημιουργεί αυτόματα (χωρίς ανθρώπινη επίβλεψη) παραδείγματα εκπαίδευσης μιας Μηχανής Διανυσμάτων Υποστήριξης (Μ.Δ.Υ., Support Vector Machine), την οποία χρησιμοποιεί κατόπιν για να κατατάσσει τμήματα των ιστοσελίδων που επιστρέφει η μηχανή αναζήτησης ως ορισμούς ή μη-ορισμούς του όρου της ερώτησης.

Στη παρούσα εργασία εξετάζουμε τρεις νέες μεθόδους αυτόματης παραγωγής παραδειγμάτων εκπαίδευσης της Μ.Δ.Υ. του παραπάνω συστήματος. Δοκιμάζουμε, επίσης, τεχνικές αφαίρεσης πλεοναζόντων υποψηφίων απαντήσεων του συστήματος. Ακόμα, μελετάμε τα κορυφαία σχετικά συστήματα που έχουν παρουσιαστεί διεθνώς και συγκρίνουμε τα αποτελέσματά τους με αυτά του δικού μας.

1.2 Διάρθρωση της εργασίας

Η εργασία είναι διαρθρωμένη ως εξής:

Στο κεφάλαιο 2 περιγράφουμε λεπτομερώς τη λειτουργία του συστήματος της εργασίας, ιδιαίτερα τις νέες μεθόδους και τεχνικές που προτείνουμε. Στο τέλος του κεφαλαίου παρουσιάζουμε και αναλύουμε τα κορυφαία σχετικά συστήματα της βιβλιογραφίας.

Στο κεφάλαιο 3 αξιολογούμε τις νέες τεχνικές και μεθόδους της εργασίας και συγκρίνουμε τα

[1] Βλ. http://trec.nist.gov

αποτελέσματα του συστήματός μας με εκείνα άλλων κορυφαίων συστημάτων.

Στο κεφάλαιο 4 γίνεται μια σύντομη ανασκόπηση των συμπερασμάτων της εργασίας και παρατίθενται προτάσεις περαιτέρω βελτίωσης του συστήματος που παρήχθη.

2. ΜΕΘΟΔΟΙ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΧΕΙΡΙΣΜΟΥ ΕΡΩΤΗΣΕΩΝ ΟΡΙΣΜΟΥ

2.1 Συστήματα ερωταποκρίσεων

Τα συστήματα ερωταποκρίσεων (question answering systems) απαντούν ερωτήματα φυσικής γλώσσας του χρήστη επιστρέφοντας τμήματα κειμένων ή γενικότερα πληροφορίες που εξάγουν από μεγάλες συλλογές κειμένων ή και ολόκληρο τον Παγκόσμιο Ιστό. Στην περίπτωση των συστημάτων ερωταποκρίσεων μεμονωμένου αποσπάσματος (single snippet), επιστρέφονται ένα ή περισσότερα αποσπάσματα κειμένου ανά ερώτηση, συνήθως μια φράση ή πρόταση το καθένα, χωρίς να γίνεται καμία προσπάθεια τα επιστρεφόμενα αποσπάσματα να συνδυαστούν, ώστε να αποτελούν μια εκτενέστερη συνολική απάντηση χωρίς πλεονασμούς (π.χ. φράσεις που επαναλαμβάνουν την ίδια πληροφορία). Στο πιο αυστηρό σενάριο χρήσης τους, τα συστήματα αυτά επιτρέπεται να επιστρέψουν μόνο ένα απόσπασμα ανά ερώτηση, οπότε η απάντησή του συστήματος θεωρείται σωστή αν η πληροφορία που ζητά η ερώτηση περιλαμβάνεται στο απόσπασμα. Σε περίπτωση που το σύστημα επιτρέπεται να επιστρέψει περισσότερα του ενός, έστω Ν, αποσπάσματα ανά ερώτηση, η ερώτηση θεωρείται πως απαντήθηκε σωστά αν η ζητούμενη πληροφορία περιλαμβάνεται σε τουλάχιστον ένα από τα επιστρεφόμενα Ν αποσπάσματα. Το σύστημα, δηλαδή, έχει Ν ευκαιρίες να βρει τη σωστή απάντηση και κάθε επιστρεφόμενο απόσπασμα θεωρείται μια αυτοτελής υποψήφια απάντηση. Αντίθετα, τα συστήματα πολλαπλών αποσπασμάτων (multi-snippet) επιχειρούν να επιστρέψουν ένα σύνολο αποσπασμάτων ανά ερώτηση, ώστε το σύνολο να αποτελεί μια εκτενή απάντηση, που να περιλαμβάνει τις σημαντικότερες σχετικές πληροφορίες αποφεύγοντας πλεονασμούς.

Υπάρχουν πολλές κατηγορίες ερωτήσεων τις οποίες καλούνται να απαντήσουν τα συστήματα ερωταποκρίσεων, όπως για παράδειγμα οι ακόλουθες:

- Ερωτήσεις με σύντομες απαντήσεις που είναι εύκολο να καθοριστούν (factual questions). Αυτές χωρίζονται περαιτέρω σε υποκατηγορίες, όπως οι ακόλουθες:
  - Ερωτήσεις προσώπου, π.χ. «Ποιος ζωγράφισε τη Μόνα Λίζα;».
  - Ερωτήσεις οργανισμού, π.χ. «Ποια εταιρία παράγει το I-Pod;».
  - Ερωτήσεις χρόνου, π.χ. «Πότε πέθανε ο Μότσαρτ;».
  - Ερωτήσεις τόπου, π.χ. «Πού βρίσκεται ο Πύργος της Πίζας;».
  - Ερωτήσεις ποσότητας, π.χ. «Πόσα χρόνια διάρκεσε ο Εκατονταετής Πόλεμος;».
  - Ερωτήσεις ορισμού, π.χ. «Τι είναι ο μυστικισμός;».
- Ερωτήσεις γνώμης (opinion questions), π.χ. «Τι θεωρείτε πως επηρέασε το αποτέλεσμα των φετινών εκλογών;»

- Ερωτήσεις περίληψης (summary questions), π.χ. «Ποια είναι η βασική ιστορία που παρουσιάζεται στο βιβλίο "Όνειρο σε Κύκλο";»

Στην παρούσα εργασία θα επικεντρωθούμε στις ερωτήσεις ορισμού. Σκοπός του συστήματος σε αυτή τη περίπτωση είναι να επιστρέψει αποσπάσματα που ορίζουν τον όρο της ερώτησης, τον οποίο καλούμε εφεξής «όρο-στόχο».

Τα πειράματα της εργασίας έγιναν στην αγγλική γλώσσα, λόγω του μεγαλύτερου αριθμού ιστοσελίδων και ηλεκτρονικών εγκυκλοπαιδειών που διατίθενται σε αυτή τη γλώσσα, αλλά οι μέθοδοι που χρησιμοποιούμε μπορούν να εφαρμοστούν και σε κείμενα άλλων γλωσσών.

2.2 Περιγραφή του συστήματος της εργασίας και των μεθόδων του

Το σύστημά μας ακολουθεί τρία βασικά στάδια επεξεργασίας: (α) την επεξεργασία της ερώτησης, (β) την εξαγωγή και επεξεργασία πιθανών απαντήσεων από μια συλλογή εγγράφων (στα πειράματά μας χρησιμοποιούμε τον Παγκόσμιο Ιστό) και (γ) την αξιολόγηση των πιθανών απαντήσεων και την επιστροφή στο χρήστη των καταλληλότερων. Εστιαζόμαστε αρχικά στην περίπτωση όπου το σύστημα επιστρέφει ένα ή περισσότερα μεμονωμένα αποσπάσματα ανά ερώτηση. Θα εξετάσουμε αργότερα πώς μπορεί να επεκταθεί, ώστε να επιστρέφει για κάθε ερώτηση ένα σύνολο αποσπασμάτων που να αποτελεί μια εκτενή απάντηση χωρίς πλεονασμούς.

Η βασική διαφορά μεταξύ εναλλακτικών συστημάτων ερωταποκρίσεων βρίσκεται στη μέθοδο με την οποία γίνεται η αξιολόγηση των πιθανών απαντήσεων στο στάδιο (γ). Στην περίπτωσή μας, στο στάδιο αυτό χρησιμοποιείται μια Μηχανή Διανυσμάτων Υποστήριξης (Μ.Δ.Υ., Support Vector Machine, SVM), μια από τις πιο επιτυχημένες μεθόδους επιβλεπόμενης μηχανικής μάθησης. Για περισσότερες πληροφορίες σχετικά με τις Μ.Δ.Υ. μπορείτε να ανατρέξετε στην εργασία του Λουκαρέλλι [Lu05] και στις εξής πηγές: [CrSh20], [CoVa95] και [Va98].

Η χρήση επιβλεπόμενης μηχανικής μάθησης απαιτεί την εκπαίδευση του συστήματος πριν τη χρήση του. Η εκπαίδευση γίνεται δίνοντας στο σύστημα ένα σύνολο από παραδείγματα εισόδων, το καθένα μαζί με την επιθυμητή απόκριση του συστήματος. Στην περίπτωσή μας, τα παραδείγματα εκπαίδευσης είναι αποσπάσματα κειμένων που έχει εξαγάγει το σύστημά μας στο στάδιο (β) για διάφορους όρους-στόχους. Κάθε απόσπασμα εκπαίδευσης πρέπει να έχει σημειωθεί ως ορισμός (αποδεκτή απάντηση της αντίστοιχης ερώτησης) ή μη-ορισμός (μη αποδεκτή απάντηση). Οι όροι που επιλέγονται για την εκπαίδευση φροντίζουμε να είναι διαφορετικοί από τους όρους που χρησιμοποιούμε κατόπιν για την αξιολόγηση του εκπαιδευμένου συστήματος.

Ένα σημαντικό μειονέκτημα της χρήσης μεθόδων επιβλεπόμενης μηχανικής μάθησης είναι ότι απαιτούν μεγάλο όγκο χειρωνακτικά ταξινομημένων παραδειγμάτων εκπαίδευσης. Για να αντιμετωπιστεί αυτό το πρόβλημα στην περίπτωση των ερωτήσεων ορισμού, οι προηγούμενες εργασίες ([Mi03], [Ga04], [Gi05], [La06]) στις οποίες βασίζεται η παρούσα έχουν προτείνει μια τεχνική

αυτόματης επισημείωσης (κατάταξης) παραδειγμάτων εκπαίδευσης. Η τεχνική αυτή κατασκευάζει αυτόματα, χωρίς ανθρώπινη επίβλεψη, μεγάλο αριθμό παραδειγμάτων εκπαίδευσης εκμεταλλευόμενη υπάρχοντες ορισμούς ηλεκτρονικών εγκυκλοπαιδειών (π.χ. Wikipedia, Encarta). Τα παραγόμενα παραδείγματα χρησιμοποιούνται για να εκπαιδευθεί η Μ.Δ.Υ. του σταδίου (γ), η οποία χρησιμοποιείται κατόπιν για να εντοπίζονται ορισμοί όρων (π.χ. ονόματα προσώπων της επικαιρότητας, νέοι τεχνικοί όροι) που δεν καλύπτονται από τις διαθέσιμες εγκυκλοπαίδειες.

Παρακάτω παρουσιάζουμε πιο λεπτομερώς τα στάδια επεξεργασίας του συστήματος της εργασίας, καθώς και τους μηχανισμούς αυτόματης εκπαίδευσής του.

2.2.1 Επεξεργασία ερώτησης και εξαγωγή πιθανών απαντήσεων

Οι ερωτήσεις που εισάγονται στο σύστημα υφίστανται προηγουμένως χειρωνακτική επεξεργασία που αφήνει σε αυτές μόνο τους όρους-στόχους, δηλαδή τους όρους των οποίων τον ορισμό καλείται να επιστρέψει το σύστημα. (Η επεξεργασία αυτή ενδέχεται να είναι δυνατόν να αυτοματοποιηθεί σε επόμενη εργασία.) Για κάθε όρο-στόχο, το σύστημα αναζητά σχετικά έγγραφα στον Παγκόσμιο Ιστό. Στην τρέχουσα υλοποίηση του συστήματος χρησιμοποιείται η μηχανή αναζήτησης Altavista[2], η οποία επιστρέφει τις πιο σχετικές ιστοσελίδες κατά φθίνουσα συνάφεια. Από αυτές, κρατάμε μόνο τις 10 πρώτες, δηλαδή τις 10 πιο σχετικές.

Σκοπός μας είναι να δείξουμε ότι το σύστημά μας καταφέρνει να βρει ορισμούς όρων οι οποίοι δεν καλύπτονται από ηλεκτρονικές εγκυκλοπαίδειες του διαδικτύου. Προκειμένου να προσομοιώσει την αναζήτηση ορισμών αυτού του είδους, το σύστημα αγνοεί κατά το στάδιο χρήσης του (μετά την εκπαίδευσή του) όσες από τις ιστοσελίδες που επιστρέφει η μηχανή αναζήτησης προέρχονται από τέτοιες εγκυκλοπαίδειες. Για το σκοπό αυτό, έχει δημιουργηθεί μια λίστα με τις διευθύνσεις των πιο γνωστών εγκυκλοπαιδειών, ώστε να αγνοούνται όσες ιστοσελίδες προέρχονται από αυτές.

Κάθε μία από τις ιστοσελίδες που απομένουν υφίσταται επεξεργασία, η οποία αφαιρεί όλα τα περιττά για τους σκοπούς μας στοιχεία από το κείμενό της (π.χ. ετικέτες HTML). Έπειτα, από το «καθαρό» κείμενο που έχει προκύψει, εξάγονται όλα τα «παράθυρα» (ακολουθίες συνεχόμενων λέξεων) μήκους 250 χαρακτήρων που περιέχουν τον όρο-στόχο στο κέντρο τους. Ο ορισμός ενός όρου είναι πιθανότερο να εμφανιστεί στις πρώτες εμφανίσεις του όρου σε ένα κείμενο. Για το λόγο αυτό, κρατάμε μόνο τα πέντε πρώτα παράθυρα από κάθε κείμενο (ιστοσελίδα). Συνολικά, λοιπόν, συλλέγουμε πέντε παράθυρα ανά κείμενο από δέκα κείμενα για κάθε όρο-στόχο, άρα πενήντα παράθυρα ανά όρο-στόχο. Κάποια από αυτά τα παράθυρα περιέχουν αποδεκτούς ορισμούς του όρου-στόχου, ενώ τα υπόλοιπα όχι.

[2] Βλ. http://www.altavista.com

2.2.2 Κατασκευή παραθύρων εκπαίδευσης

Αυτό το στάδιο συμβαίνει μόνο κατά τη λειτουργία της εκπαίδευσης. Τα παράθυρα εκπαίδευσης που έχουμε συλλέξει για διάφορους όρους-στόχους πρέπει τώρα να μετατραπούν στη διανυσματική μορφή που απαιτεί η Μ.Δ.Υ. και να δοθούν σε αυτή για την εκπαίδευσή της, αφού πρώτα επισημειωθούν με την ορθή τους κατηγορία (ορισμοί ή μη-ορισμοί). Η επισημείωση των παραθύρων εκπαίδευσης αποτελεί ένα από τα βασικά προβλήματα που έχουμε να αντιμετωπίσουμε κατά την εκπαίδευση. Η χειρωνακτική επισημείωσή τους είναι χρονοβόρα και έτσι περιορίζει και το πλήθος των παραθύρων εκπαίδευσης που είναι δυνατόν να χρησιμοποιηθούν. Για αυτό, όπως προαναφέρθηκε, έχουν προταθεί σε προηγούμενες εργασίες μέθοδοι αυτόματης επισημείωσης των παραθύρων εκπαίδευσης.

Η βασική ιδέα είναι απλή. Κατά την εκπαίδευση του συστήματος, χρησιμοποιούμε όρους-στόχους για τους οποίους διαθέτουμε ορισμούς από εγκυκλοπαίδειες, αντίθετα από ό,τι συμβαίνει κατά τη χρήση του εκπαιδευμένου συστήματος. Ας υποθέσουμε ότι έχουμε στην διάθεσή μας έναν τέτοιο όρο-στόχο, καθώς και ένα παράθυρο κειμένου που τον περιέχει, το οποίο έχει εξαχθεί από ιστοσελίδες του διαδικτύου με τον τρόπο που περιγράψαμε παραπάνω. Μπορούμε να υπολογίσουμε με κάποιο μέτρο την ομοιότητα μεταξύ του παραθύρου και του αντίστοιχου ορισμού που διαθέτουμε από εγκυκλοπαίδειες και να εκτιμήσουμε αν το παράθυρο είναι αποδεκτός ορισμός με βάση την ομοιότητα. Με τον τρόπο αυτό μπορούμε να επισημειώσουμε το παράθυρο ως ορισμό ή μη ορισμό και επομένως να δημιουργήσουμε από το παράθυρο ένα θετικό ή αρνητικό παράδειγμα εκπαίδευσης. Ομοίως, μπορούμε να δημιουργήσουμε και πολλά άλλα (θετικά και αρνητικά) παραδείγματα εκπαίδευσης. Κατόπιν μπορούμε να εκπαιδεύσουμε τη Μ.Δ.Υ. στα παραδείγματα εκπαίδευσης, ώστε να είναι σε θέση κατόπιν να κατατάσσει στις δύο κατηγορίες (ορισμούς και μη ορισμούς) παράθυρα όρων-στόχων για τους οποίους δεν διαθέτουμε ορισμούς από εγκυκλοπαίδειες.

Ένα απλό μέτρο ομοιότητας είναι να μετρηθούν οι κοινές λέξεις του παραθύρου και του αντίστοιχου ορισμού της εγκυκλοπαίδειας[3]. Παράδειγμα χρήσης αυτού του μέτρου είναι το παρακάτω.

[3] Στα πειράματά μας, λαμβάνουμε ορισμούς εγκυκλοπαίδειας χρησιμοποιώντας τη λειτουργία «define» της μηχανής αναζήτησης Google, η οποία επιστρέφει ορισμούς από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια. Βλ. http://www.google.com/help/features.html#definitions.

Όρος-στόχος:

Archimedes

Παράθυρο κειμένου από ιστοσελίδα:

nova | infinite secrets | library resource kit | who was archimedes? | pbs who was **archimedes**?

by [author] infinite secrets homepage **archimedes** of syracuse was one of the greatest

mathematicians in history.

Ορισμός εγκυκλοπαίδειας:

A Greek mathematician living from approximately 287 BC to 212 BC in Syracuse. He

invented much plane geometry, studying the circle, parabola and three-dimensional geometry

of the sphere as well as studying physics. See also Archimedean solid.

Κοινές λέξεις:

of, the, in, Syracuse

Παρατηρούμε πως παρότι και τα δύο κείμενα περιέχουν ορισμούς του όρου-στόχου «Archimedes», οι κοινές λέξεις τους είναι λίγες και δεν φανερώνουν πραγματική ομοιότητα, αφού θα μπορούσαν να είναι κοινές μεταξύ οποιωνδήποτε κειμένων που αναφέρονται στον Αρχιμήδη. Το μέτρο αποτυγχάνει, επίσης, να εντοπίσει τη λέξη «mathematician» ως κοινή, λόγω της διαφορετικής της κατάληξης σε καθένα από τα κείμενα.

Ένα άλλο πρόβλημα είναι πως ένας ορισμός μπορεί να διατυπωθεί με πολλούς τρόπους. Συγκρίνοντας το παράθυρο με ένα μόνο ορισμό εγκυκλοπαίδειας περιοριζόμαστε μόνο στις κοινές λέξεις που έχει το παράθυρο με τη συγκεκριμένη διατύπωση του ορισμού. Το πρόβλημα αυτό είναι ακόμα εντονότερο όταν οι όροι-στόχοι έχουν παραπάνω από μία δυνατές σημασίες. Για να γίνει πιο κατανοητό το πρόβλημα, παρουσιάζουμε παρακάτω διαφορετικούς ορισμούς του όρου «Γαλαξίας», που λάβαμε από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια του διαδικτύου.

Όρος-στόχος:

Galaxy

Ορισμοί από το εγκυκλοπαίδειες και γλωσσάρια του διαδικτύου:

A large aggregation of stars, bound together by gravity. There are three major classifications of galaxies-spiral, elliptical, and irregular.

a very large cluster of stars (tens of millions to trillions of stars) gravitationally bound together.

an organized system of many hundreds of millions of stars, often mixed with gas and dust. The universe contains billions of galaxies.

a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.

A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant that their light takes millions of years to reach the Earth.

2.2.2.1 Η μέθοδος του Γαλάνη

Τα προβλήματα αυτά εντόπισε και αντιμετώπισε ο Γαλάνης [Ga04] με τον παρακάτω αλγόριθμο, ο οποίος υπολογίζει την ομοιότητα μεταξύ ενός παραθύρου (παραδείγματος εκπαίδευσης) ενός όρου-στόχου και ενός συνόλου ορισμών για τον ίδιο όρο-στόχο που έχουμε στη διάθεσή μας από εγκυκλοπαίδειες.

Ο αλγόριθμος του Γαλάνη εκτελεί αρχικά τα εξής βήματα προεπεξεργασίας:

- Αφαίρεση από το παράθυρο και τους ορισμούς εγκυκλοπαιδειών των 100 συχνότερων λέξεων που εμφανίζονται σε αγγλικά κείμενα (π.χ. “the”, “be”, “of”, “and”, “a”, “in”, “to”, “have”, “it”, “to”, “for”, “i”, “that”, “you”, “he”, “on”, “with”, “do”, “at”, “by”, “not”, “this”).

  Οι 100 συχνότερες λέξεις έχουν προκύψει από το British National Corpus[4]. Η αφαίρεση αυτών των λέξεων γίνεται διότι κατά τη σύγκριση δύο κειμένων, η εύρεση κοινών λέξεων που είναι πολύ συχνές δεν φανερώνει ομοιότητα.

[4]Βλ. http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bncreadme.html

- Εφαρμογή ενός αλγορίθμου που αποκόπτει την κατάληξη κάθε λέξης αφήνοντας μόνο τη ρίζα της (stemmer, π.χ. το «invented» γίνεται «invent»). Ο αλγόριθμος αποκοπής που χρησιμοποιήθηκε είναι εκείνος του Porter[5].

- Διαγραφή από κάθε παράθυρο των ειδικών συμβόλων (!@&^%$#., κλπ.).

Στη συνέχεια υπολογίζεται για κάθε παράθυρο ένας αριθμός (score). Αυτός ο αριθμός προκύπτει από τη σύγκριση του παραθύρου με όλους τους ορισμούς του όρου-στόχου που έχουμε στην διάθεσή μας και δείχνει την ομοιότητα του παραθύρου με τους ορισμούς. Συγκεκριμένα, όσο μεγαλύτερος ο αριθμός τόσο μεγαλύτερη η ομοιότητα και άρα τόσο μεγαλύτερη και η πιθανότητα το παράθυρο να είναι πράγματι ορισμός. Στη συνέχεια περιγράφεται λεπτομερέστερα ο τρόπος υπολογισμού αυτού του αριθμού (score).

Σε κάθε λέξη του παραθύρου δίνουμε ένα βάρος w, το οποίο υπολογίζουμε ως εξής:

$$w = fdef * idf$$

Όπου:

w: το βάρος της λέξης

fdef:το ποσοστό των διαθέσιμων ορισμών που περιέχουν την λέξη

idf (inverse document frequency): η αντίστροφη συχνότητα εγγράφων της λέξης.

Το idf ορίζεται ως εξής:

$$idf = 1 + \log\left(N / df\right)$$

Όπου:

N: ο ολικός αριθμός των εγγράφων του British National Corpus (BNC)

df: ο αριθμός των εγγράφων του British National Corpus που περιέχουν τη λέξη.

Παρατηρούμε ότι δεν έχει το ίδιο βάρος κάθε λέξη του παραθύρου. Αν η λέξη εμφανίζεται σε μεγάλο ποσοστό των ορισμών (υψηλό fdef), υπάρχει μεγαλύτερη πιθανότητα η εμφάνισή της σε κάποιο παράθυρο να φανερώνει ότι είναι και αυτό ορισμός. Επίσης, όσο πιο σπάνια είναι μια λέξη (υψηλό idf), τόσο πιο απίθανο είναι η εμφάνισή της και στον ορισμό και στο παράθυρο να οφείλεται σε σύμπτωση.

[5] Βλ. http://www.tartarus.org/~martin/PorterStemmer

Τελικά το score κάθε παραθύρου υπολογίζεται από τον παρακάτω τύπο.

$$score = \frac{\sum_{i=1}^{n} w_i}{n}$$

Όπου:

n: ο αριθμός των λέξεων του παραθύρου. Λέξεις που εμφανίζονται στο παράθυρο πολλές φορές υπολογίζονται μόνο μία φορά.

w_i: το βάρος της λέξης i.

Παράθυρα εκπαίδευσης των οποίων το score υπερβαίνει ένα άνω κατώφλι σημειώνονται ως ορισμοί, ενώ παράθυρα εκπαίδευσης των οποίων το score είναι μικρότερο ενός κάτω κατωφλίου σημειώνονται ως μη ορισμοί. Τα παράθυρα εκπαίδευσης των οποίων το score βρίσκεται μεταξύ των δύο κατωφλίων αγνοούνται κατά την εκπαίδευση της Μ.Δ.Υ., γιατί δεν είμαστε επαρκώς σίγουροι για την ορθή κατηγορία τους. Η χρήση και ο τρόπος επιλογής των δύο κατωφλίων επεξηγούνται στη συνέχεια, μετά την παρουσίαση της επέκτασης της μεθόδου του Γαλάνη που πρότεινε ο Γιακουμής.

2.2.2.2 Μέθοδος του Γιακουμή

Ενώ η μέθοδος του Γαλάνη συγκρίνει μεμονωμένες λέξεις του παραθύρου και των διαθέσιμων ορισμών, ο Γιακουμής [Gi05] πρόσθεσε και την σύγκριση ν-γραμμάτων (n-gram) λέξεων, δηλαδή ακολουθιών ν συνεχόμενων λέξεων. Οι συγκρίσεις που γίνονται εξαρτώνται από την τιμή της παραμέτρου ν. Στην περίπτωση του ν = 1, η νέα μέθοδος κάνει πάλι συγκρίσεις μεταξύ μεμονωμένων λέξεων και είναι παρόμοια με την αρχική μέθοδο του Γαλάνη, αλλά έχει το πλεονέκτημα ότι επιστρέφει μια κανονικοποιημένη τιμή στο [0, 1]. Ουσιαστική διαφορά υπάρχει για ν > 1. Για παράδειγμα, όταν ν = 3 η νέα μέθοδος συγκρίνει όλες τις μεμονωμένες λέξεις, όλες τις ακολουθίες λέξεων μήκους 2 και όλες τις ακολουθίες λέξεων μήκους 3.

Στη νέα μέθοδο του Γιακουμή, υπολογίζουμε αρχικά το βάρος κάθε ν-γράμματος του παραθύρου. Το βάρος κάθε ν-γράμματος (περιλαμβανόμενων και τον μονογραμμάτων, δηλαδή λέξεων) ορίζεται ως εξής:

$$w = fdef * avgidf$$

Όπου:

w: το βάρος του ν-γράμματος

fdef: το ποσοστό των διαθέσιμων ορισμών που περιέχουν το ν-γραμμα

avgidf: ο μέσος όρος των idf των λέξεων του ν-γράμματος.

Ο υπολογισμός του score του παραθύρου γίνεται τώρα ως εξής:

$$score = \frac{\sum_{n=1}^{m} \frac{\sum_{\gamma \in grams_C(n) \cap \gamma \in grams_W(n)} w_\gamma}{\sum_{\gamma \in grams_C(n)} w_\gamma}}{m}$$

Όπου:

m: το μέγιστο μήκος n-γραμμάτων που εξετάζουμε

γ: ένα n-γραμμα

$grams_C(n)$: το σύνολο των n-γραμμάτων (μήκους n) των διαθέσιμων ορισμών

$grams_W(n)$: το σύνολο των n-γραμμάτων (μήκους n) του παραθύρου

***w_γ*:** το βάρος του n-γράμματος γ.

Τελικά, είτε χρησιμοποιούμε την αρχική μέθοδο του Γαλάνη είτε την επέκτασή της του Γιακουμή, σε κάθε παράθυρο εκπαίδευσης του όρου-στόχου θα έχει δοθεί ένα score. Μας μένει να σημειώσουμε τα παράθυρα εκπαίδευσης ως ορισμούς ή μη-ορισμούς με βάση αυτό το score. Για το σκοπό αυτό, χρησιμοποιούμε (όπως και στις εργασίες των Γαλάνη και Γιακουμή) δύο κατώφλια t- (κάτω κατώφλι) και t+ (άνω κατώφλι), τα οποία επιλέγονται έτσι ώστε να ισχύουν τα εξής:

- Τα παράθυρα εκπαίδευσης που έχουν score μεγαλύτερο του άνω κατωφλιού να είναι παράθυρα ορισμού με μεγάλη πιθανότητα.
- Τα παράθυρα εκπαίδευσης που έχουν score μικρότερο του κάτω κατωφλιού να είναι παράθυρα μη-ορισμού με μεγάλη πιθανότητα.

Τα δύο αυτά κατώφλια χωρίζουν τα διαθέσιμα παράθυρα (υποψήφια παραδείγματα εκπαίδευσης) σε τρεις κατηγορίες. Πρώτον, σε παράθυρα για τα οποία είμαστε σχεδόν βέβαιοι ότι είναι ορισμοί (score > t_+). Δεύτερον, σε παράθυρα για τα οποία είμαστε σχεδόν βέβαιοι ότι δεν είναι

ορισμοί (score < t_-). Τρίτον, σε παράθυρα για τα οποία δεν μπορούμε να αποφασίσουμε με βεβαιότητα για την κατηγορία τους ($t_- \leq$ score $\leq t_+$). Για την εκπαίδευση της Μ.Δ.Υ. χρησιμοποιούμε μόνο τις δύο πρώτες κατηγορίες, ενώ τα παράθυρα της τρίτης αγνοούνται.

Αν πολλά υποψήφια παράθυρα εκπαίδευσης καταταγούν στην τρίτη κατηγορία, τότε είναι πιθανόν να μείνουν πολύ λίγα παράθυρα, που να μην επαρκούν για την εκπαίδευση της Μ.Δ.Υ. Σε μια τέτοια περίπτωση, πρέπει να χρησιμοποιηθούν περισσότερες ερωτήσεις/κείμενα εκπαίδευσης, ώστε ο αριθμός των παραθύρων των δύο πρώτων κατηγοριών να αυξηθεί. Εναλλακτικά, μπορεί κανείς να μειώσει το άνω κατώφλι και να αυξήσει το κάτω, ώστε να αγνοούνται λιγότερα παράθυρα εκπαίδευσης, διακινδυνεύοντας όμως να αυξηθεί ο αριθμός των παραθύρων εκπαίδευσης που σημειώνονται λάθος. Επίσης, προσοχή πρέπει να δοθεί στην ισορροπία μεταξύ του πλήθους των παραθύρων που κατατάσσονται στην πρώτη κατηγορία και τη δεύτερη. Αν μια κατηγορία αποκτήσει πολύ περισσότερα παραδείγματα εκπαίδευσης από την άλλη, υπάρχει ο κίνδυνος η Μ.Δ.Υ. να μάθει να ταξινομεί όλα τα παράθυρα σε εκείνη.

2.2.2.3 Μέθοδος του κεντροειδούς

Σε αυτή τη μέθοδο, που προτάθηκε από τους Cui κ.ά. [Cu06], κατασκευάζουμε ένα «κεντροειδές» από όλα τα παράθυρα του όρου-στόχου που εξήχθησαν από τη συλλογή των κειμένων. Στη συνέχεια, σε κάθε παράθυρο αντιστοιχίζεται ένα score, ανάλογα με την ομοιότητά του με το κεντροειδές.

Πιο συγκεκριμένα, αρχικά υπολογίζουμε το βάρος της κάθε λέξης που εμφανίζεται στα παράθυρα του όρου-στόχου, χρησιμοποιώντας τον παρακάτω τύπο. Πριν γίνουν οι υπολογισμοί αφαιρούνται από τα παράθυρα οι 100 πιο συχνές λέξεις και οι καταλήξεις τους (stemming).

$$w = -\log\left(\frac{\mathrm{Co}_{\mathrm{term}\cap\mathrm{word}}}{\mathrm{Sf}_{\mathrm{term}} + \mathrm{Sf}_{\mathrm{word}}}\right) * idf$$

Όπου:

w: το βάρος της λέξης

$\mathrm{Co}_{\mathrm{term}\cap\mathrm{word}}$: το πλήθος των παραθύρων του όρου-στόχου που περιλαμβάνουν και τη λέξη της οποίας υπολογίζουμε το βάρος

$\mathrm{Sf}_{\mathrm{term}}$: το πλήθος των παραθύρων του όρου-στόχου

$\mathrm{Sf}_{\mathrm{word}}$: το πλήθος των παραθύρων της λέξης της οποίας υπολογίζουμε το βάρος

idf (inverse document frequency): η αντίστροφη συχνότητα εγγράφων της λέξης της οποίας υπολογίζουμε το βάρος.

Οι παραπάνω υπολογισμοί γίνονται όπως στην εργασία των Cui κ.ά., με τη διαφορά ότι χρησιμοποιούμε παράθυρα του όρου-στόχου, ενώ οι Cui κ.ά. χρησιμοποιούν προτάσεις. Ο συμβολισμός sf (sentence frequency) προέρχεται από την εργασία των Cui κ.ά. Ο παραπάνω τύπος βασίζεται στον τύπο της αμοιβαίας πληροφορίας (mutual information), στον οποίο έχει προστεθεί ο παράγοντας idf.

Οι λέξεις που έχουν βάρος μεγαλύτερο από το μέσο όρο των βαρών συν την τυπική απόκλιση περιλαμβάνονται στο κεντροειδές του όρου-στόχου. Στη συνέχεια, το κεντροειδές και κάθε ένα από τα παράθυρα μετατρέπονται σε δυαδικά διανύσματα, όλα τόσων διαστάσεων όσες είναι συνολικά οι διαφορετικές λέξεις που εμφανίζονται στο κεντροειδές ή τα παράθυρα. Κάθε συντεταγμένη των διανυσμάτων (0 ή 1) δείχνει αν το κεντροειδές ή παράθυρο, αντίστοιχα, περιλαμβάνει ή όχι τη λέξη που αντιστοιχεί στη συντεταγμένη. Κατόπιν υπολογίζεται για κάθε παράθυρο η ομοιότητά του με το κεντροειδές, χρησιμοποιώντας την ομοιότητα συνημίτονου (cosine similarity):

$$\mathrm{score} = \frac{\mathrm{vector}_{\text{κεντροειδούς}} * \mathrm{vector}_{\text{παραθύρου}}}{\left|\mathrm{vector}_{\text{κεντροειδούς}}\right| + \left|\mathrm{vector}_{\text{παραθύρου}}\right|}$$

Όπου:

vector: το διάνυσμα του κεντροειδούς ή παραθύρου

* το εσωτερικό γινόμενο,

|vector|: το μέτρο του διανύσματος.

Όπως και στις μεθόδους των Γαλάνη και Γιακουμή, το παραπάνω score μπορεί να χρησιμοποιηθεί για να διαχωριστούν, μέσω δύο κατωφλίων, τα παράθυρα σε τρεις κατηγορίες: παράθυρα που είναι ορισμοί με μεγάλη βεβαιότητα (υψηλό score), παράθυρα που με μεγάλη βεβαιότητα δεν είναι ορισμοί (χαμηλό score) και παράθυρα για τα οποία έχουμε μεγάλη αβεβαιότητα αν είναι ή όχι ορισμοί.

2.2.2.4 Αυτόματη επισημείωση παραδειγμάτων με το ROUGE

Το πακέτο ROUGE [Li04], που χρησιμοποιείται συχνά για τη σύγκριση χειρωνακτικά κατασκευασμένων περιλήψεων με περιλήψεις που έχουν παραχθεί από συστήματα, παρέχει ένα σύνολο μέτρων για τον υπολογισμό της ομοιότητας μεταξύ κειμένων. Συγκεκριμένα, το μέτρο που θα χρησιμοποιήσουμε για την αυτόματη επισημείωση των παραδειγμάτων είναι το ROUGE-W. Το ROUGE-W λειτουργεί βρίσκοντας πρώτα τη μέγιστη κοινή υποακολουθία λέξεων ανάμεσα σε δύο κείμενα (ή μια από τις μέγιστες, αν υπάρχουν πολλές), αφού πρώτα αφαιρέσει τις καταλήξεις των λέξεων. Για παράδειγμα για τις δύο ακολουθίες λέξεων:

X: [A B C D E F G]

Y: [A B E C G F D]

η μέγιστη κοινή υποακολουθία είναι οι [A B C D] και [A B E F].

Το ROUGE-W βαθμολογεί υψηλότερα (μεγαλύτερη ομοιότητα) ζευγάρια κειμένων, όταν μεγάλα τμήματα της μέγιστης κοινής υποακολουθίας εμφανίζονται αυτούσια (χωρίς να παρεμβαίνουν άλλες λέξεις) και στα δύο κείμενα. Στο παραπάνω παράδειγμα, το μόνο τέτοιο τμήμα μήκους 2 είναι το [A B]. Μπορούμε να θεωρήσουμε ότι η μέγιστη κοινή υποακολουθία [A B C D] του παραδείγματος διαμερίζεται σε τμήματα που εμφανίζονται αυτούσια και στα δύο κείμενα ως εξής: [A B | C | D], δηλαδή έχουμε τρία τμήματα, μήκους 2, 1 και 1. Ομοίως διαμερίζεται η μέγιστη κοινή υποακολουθία κάθε ζεύγους κειμένων X και Y. Η ομοιότητα των X και Y υπολογίζεται με τον παρακάτω τύπο, όπου WLCS σημαίνει «Weighted Longest Common Subsequence»:

$$WLCS(X,Y) = \sum_{k=1}^{m} f(s_k)$$

Όπου:

m: το πλήθος των τμημάτων της διαμέρισης της μέγιστης κοινής υποακολουθίας των X και Y

s_k: το μήκος του k-στού τμήματος

$f(s_k)$: μια συνάρτηση βάρους, που δίνει μεγαλύτερο βάρος σε μεγάλα τμήματα.

Το ROUGE-W υπολογίζεται με τους παρακάτω τύπους:

$$R = f^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right)$$

$$P = f^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right)$$

$$F = \frac{(1+\beta^2) R\, P}{R + \beta^2 P}$$

Όπου:

f^{-1}: η αντίστροφη συνάρτηση της f

m, n: τα μήκη σε λέξεις των κειμένων X και Y αντίστοιχα

β: παράμετρος του F-measure. Για β=1, δίδεται ίσο βάρος στο R και το P. Για β=2, δίδεται διπλάσιο βάρος στο R από ό,τι στο P.

Χονδρικά, το R μετρά τι μέρος του X αποτελείται από τμήματα της μέγιστης κοινής υποακολουθίας· αντίστοιχα για το Y. Στην περίπτωσή μας, αφού υπολογιστεί η ομοιότητα μεταξύ του παραθύρου προς επισημείωση (που θεωρούμε ότι είναι το X) και κάθε διαθέσιμου ορισμού εγκυκλοπαίδειας (που θεωρούμε ότι είναι το Y), η τελική τιμή του ROUGE-W είναι η μέγιστη αυτών. Ως συνάρτηση βάρους στο πακέτο ROUGE υλοποιείται η $f(\mathbf{s_k}) = \mathbf{s_k}^{\alpha}$, όπου α παράμετρος βάρους (α > 1). Οι Lin κ.ά. [Li04] δίνουν στο β μια σχετικά μεγάλη τιμή (β = 8), δηλαδή δίνουν μεγαλύτερο βάρος στο R, ώστε να θεωρείται σημαντικότερο το παράθυρο (X) να αποτελείται σε μεγάλο βαθμό από τμήματα της μέγιστης κοινής υποακολουθίας.

2.2.2.5 Αυτόματη επισημείωση παραδειγμάτων με δεύτερη, βοηθητική Μ.Δ.Υ.

Μια άλλη προσέγγιση που δοκιμάσαμε είναι να χρησιμοποιηθούν κατά την επισημείωση των παραδειγμάτων εκπαίδευσης πολλά μέτρα ομοιότητας, τα οποία συνδυάζονται με τη χρήση μιας δεύτερης, βοηθητικής Μ.Δ.Υ., παρόμοιας με εκείνη που χρησιμοποιήθηκε από τους Μαλακασιώτη και Ανδρουτσόπουλο [MaAn07] στο πρόβλημα του εντοπισμού παραφράσεων.

Στην περίπτωση αυτή, συγκρίνουμε το κάθε παράθυρο εκπαίδευσης (της κύριας Μ.Δ.Υ.) με τους διαθέσιμους ορισμούς (του ιδίου όρου) που έχουμε από εγκυκλοπαίδειες, χρησιμοποιώντας περισσότερα του ενός μέτρα ομοιότητας. Κατόπιν χρησιμοποιούμε τα αποτελέσματα των συγκρίσεων (μία τιμή για κάθε μέτρο) ως συνιστώσες ενός διανύσματος. Λαμβάνουμε έτσι ένα διάνυσμα για κάθε παράθυρο εκπαίδευσης. Στη συνέχεια, κατατάσσουμε χειρωνακτικά ορισμένα από αυτά τα διανύσματα, ανάλογα με το αν αντιστοιχούν σε παράθυρα ορισμών ή όχι, και τα χρησιμοποιούμε για να εκπαιδεύσουμε τη βοηθητική Μ.Δ.Υ. Η βοηθητική Μ.Δ.Υ. μαθαίνει έτσι να κατατάσσει (ως ορισμούς ή μη ορισμούς) παράθυρα για τα οποία διαθέτουμε ορισμούς του ιδίου όρου-στόχου από εγκυκλοπαίδειες, εξετάζοντας με πολλά μέτρα πόσο πολύ μοιάζουν τα παράθυρα με τους ορισμούς τους. Κατόπιν, χρησιμοποιούμε τη βοηθητική Μ.Δ.Υ. για να κατατάξουμε (να σημειώσουμε) αυτόματα πολύ περισσότερα παράθυρα αυτού του είδους (παράθυρα που συνοδεύονται από ορισμούς εγκυκλοπαιδειών). Τα καταταγμένα (σημειωμένα) παράθυρα που προκύπτουν χρησιμοποιούνται ως παραδείγματα εκπαίδευσης της κύριας Μ.Δ.Υ., που μαθαίνει να κατατάσσει παραδείγματα σε ορισμούς ή μη ορισμούς χωρίς να εξετάζει εγκυκλοπαίδειες (αφού χρησιμοποιείται για να βρει ορισμούς όρων που δεν περιλαμβάνονται σε εγκυκλοπαίδειες).

Τα μέτρα ομοιότητας που χρησιμοποιήσαμε μπορούν να εφαρμοστούν τόσο σε δύο συμβολοσειρές s_1 και s_2, όσο και σε άλλα ζεύγη συμβολοσειρών που προκύπτουν από τις s_1 και s_2. Τα ζεύγη που χρησιμοποιήσαμε είναι τα ακόλουθα (πρόκειται για υποσύνολο των ζευγών της εργασίας του Μαλακασιώτη):

Ζευγάρι 1: Οι αυθεντικές συμβολοσειρές s_1, s_2.

Ζευγάρι 2: Οι συμβολοσειρές s_1, s_2, αφού αφαιρεθούν από τις λέξεις που περιλαμβάνονται στις συμβολοσειρές οι καταλήξεις.

Ζευγάρι 3: Οι συμβολοσειρές s_1, s_2, αφού αφαιρεθούν από αυτές όλες οι λέξεις που δεν είναι ουσιαστικά.

Ζευγάρι 4: Όπως το προηγούμενο ζευγάρι, αλλά αφαιρούνται και οι καταλήξεις των ουσιαστικών (stemming).

Ζευγάρι 5: Οι συμβολοσειρές s_1, s_2, αφού αφαιρεθούν από αυτές όλες οι λέξεις που δεν είναι ρήματα.

Ζευγάρι 6: Όπως το προηγούμενο ζευγάρι, αλλά αφαιρούνται και οι καταλήξεις των ρημάτων.

Τα μέτρα ομοιότητας που χρησιμοποιήσαμε, τα οποία εφαρμόζονται σε κάθε ένα από τα παραπάνω ζεύγη, είναι τα ακόλουθα. Το πρώτο εξετάζει τους χαρακτήρες των συμβολοσειρών των ζευγών, ενώ τα υπόλοιπα τις λέξεις τους. Οι ορισμοί των μέτρων δίνονται στην εργασία των Μαλακασιώτη και Ανδρουτσόπουλου [MaAn07].

- Απόσταση Jaro-Winkler
- Απόσταση Manhattan
- Ευκλείδεια απόσταση
- Ομοιότητα συνημίτονου
- Συντελεστής ταιριάσματος (Matching coefficient)
- Συντελεστής Dice
- Συντελεστής Jaccard

Τα περισσότερα από τα παραπάνω μέτρα έχουν και παραλλαγές που συγκρίνουν ν-γράμματα (ακολουθίες) λέξεων ή χαρακτήρων, αντί για μεμονωμένες λέξεις ή χαρακτήρες. Έτσι ο αριθμός των μέτρων ουσιαστικά διπλασιάζεται (για συγκεκριμένο ν).

Ένα πρόβλημα που μένει να αντιμετωπίσουμε είναι ότι τα παραπάνω μέτρα συγκρίνουν δύο μόνο συμβολοσειρές μεταξύ τους. Στην περίπτωσή μας, όμως, έχουμε να συγκρίνουμε μία συμβολοσειρά (το παράθυρο, παράδειγμα εκπαίδευσης) με ένα σύνολο συμβολοσειρών (τους διαθέσιμους ορισμούς εγκυκλοπαιδειών). Προκειμένου να αντιμετωπίσουμε αυτό το πρόβλημα, για κάθε μέτρο λαμβάνουμε δύο αποτελέσματα, m_1 (μέση ομοιότητα) και m_2 (μέγιστη ομοιότητα), όπως αυτά ορίζονται ακολούθως.

$$m_1(win, defs) = \frac{\sum_{i=1}^{n} f(win, defs_i)}{n}$$

$$m_2(win, defs) = \max\left(f(win, defs_i)\right)$$

Όπου:

win: το παράθυρο ή μια συμβολοσειρά που προκύπτει από αυτό

defs: το σύνολο των διαθέσιμων ορισμών ή των συμβολοσειρών που προκύπτουν από αυτούς

defs$_i$: ο i-στός διαθέσιμος ορισμός ή η συμβολοσειρά που προκύπτει από αυτόν

n: το πλήθος των διαθέσιμων ορισμών από εγκυκλοπαίδειες

f: ένα από τα μέτρα ομοιότητας.

2.2.3 Αναπαράσταση των παραθύρων ως διανύσματα

Όπως προαναφέρθηκε, για κάθε ερώτηση που δίνεται στο σύστημα (κατά την εκπαίδευση ή τη χρήση του) συλλέγουμε σελίδες από το διαδίκτυο και από αυτές εξάγουμε τα παράθυρα που περιέχουν τον όρο-στόχο της ερώτησης. Κάθε παράθυρο θα πρέπει σε αυτό το στάδιο να μετατραπεί σε ένα διάνυσμα, ώστε είτε να προστεθεί στα δεδομένα εκπαίδευσης της (κύριας) Μ.Δ.Υ. (κατά την εκπαίδευσή της) είτε να ρωτηθεί η Μ.Δ.Υ. για την κατηγορία του (ορισμός ή μη ορισμός, κατά τη χρήση της εκπαιδευμένης Μ.Δ.Υ.).

Κάθε παράθυρο παριστάνεται ως ένα διάνυσμα που αποτελείται από χαρακτηριστικά (features), δηλαδή τιμές ιδιοτήτων (attributes). Οι πρώτες 22 ιδιότητες έχουν επιλεγεί χειρωνακτικά από τη Μηλιαράκη [Mi03], βάσει πειραμάτων σε δεδομένα των διαγωνισμών TREC.

2.2.3.1 Χειρωνακτικά επιλεγμένες ιδιότητες

Οι 22 χειρωνακτικά επιλεγμένες ιδιότητες είναι οι εξής. Οι πρώτες τρεις είναι αριθμητικές ιδιότητες με ακέραιες τιμές. Οι υπόλοιπες είναι δυαδικές και δείχνουν η κάθε μία αν το παράθυρο περιέχει (τιμή 1) ή όχι (τιμή 0) μία συγκεκριμένη φράση.

1. Η **κατάταξη (ranking) του κειμένου** από το οποίο προέρχεται το παράθυρο, δηλαδή η σειρά με την οποία επέστρεψε τη σελίδα η μηχανή αναζήτησης.

Έχει παρατηρηθεί ότι συνήθως οι ζητούμενοι ορισμοί βρίσκονται στα πρώτα κείμενα που επιστρέφονται παρά στα τελευταία.

2. Η **θέση του παραθύρου** μέσα στο έγγραφο, δηλαδή αν πρόκειται για την πρώτη, δεύτερη κτλ. εμφάνιση του όρου στο κείμενο.

   Είναι συνηθέστερο ένας όρος να ορίζεται στην αρχή ενός κειμένου.

3. Το **πλήθος των κοινών λέξεων του παραθύρου.**

   Τα παράθυρα ορισμού ενός όρου-στόχου έχουν συνήθως κοινές λέξεις μεταξύ τους. Βρίσκοντας τις κοινές λέξεις όλων των παραθύρων του όρου-στόχου που περιέχονται στα έγγραφα που επέστρεψε η μηχανή αναζήτησης, δημιουργούμε ένα κεντροειδές. Αυτό είναι παρόμοιο με εκείνο της ενότητας 2.2.2.3, αλλά οι λέξεις του κεντροειδούς επιλέγονται τώρα απλά βάσει της συχνότητάς τους. Στα πειράματά μας, το κεντροειδές περιείχε τις 20 συχνότερες λέξεις των παραθύρων του όρου-στόχου. Όσο λιγότερο απέχει ένα παράθυρο από το κεντροειδές, με άλλα λόγια όσο περισσότερες από τις κοινές λέξεις έχει, τόσο μεγαλύτερη η πιθανότητα να είναι ορισμός.

   Κατά τον υπολογισμό του κεντροειδούς αφαιρούνται και πάλι οι 100 πιο συχνές λέξεις της αγγλικής γλώσσας.

4. Η φράση **«such … as *όρος-στόχος*»**

   Παράδειγμα : *«such antibiotics as amoxicillin»*

5. Η φράση **«*όρος-στόχος* and other»**

   Παράδειγμα : «*broken bones and other injuries»*

6. Η φράση **«*όρος-στόχος* or other»**

   Παράδειγμα : *«cats or other animals»*

7. Η φράση **«especially *όρος-στόχος*»**

   Παράδειγμα : *«some plastics especially Teflon»*

8. Η φράση **«including *όρος-στόχος*»**

   Παράδειγμα : *«some amphibians including frog»*

9. **Παρενθέσεις μετά** τον όρο-στόχο

   Παράδειγμα : *«sodium chloride (salt)»*

10. **Παρενθέσεις πριν** τον όρο-στόχο

    Παράδειγμα : *«(Vitamin B1) thiamine»*

11. Η φράση **«*όρος-στόχος* is a»**

    Ακριβέστερα αναζητείται μια φράση της μορφής «*όρος-στόχος* is/are/was/were a/an/the»

    Παράδειγμα : *«Galileo was a great astronomer»*

12. **Κόμμα μετά** τον όρο-στόχο

    Παράδειγμα : *«amoxicillin, an antibiotic»*

13. Η φράση **«*όρος-στόχος* which is/was/are/were»**

    Παράδειγμα : *«tsunami which is a giant wave»*

14. Η φράση **«*όρος-στόχος* like»**

    Παράδειγμα : *«antibiotics like amoxicillin»*

15. Η φράση **«*όρος-στόχος*, …, is/was/are/were»**

    Παράδειγμα : *«amphibians, like frogs, are animals that can live both on land and in water»*

16. Η φράση **«*όρος* or»**

    Παράδειγμα : *«autism or some other type of disorder»*

17. Ένα από τα ρήματα **«can», «refer», «have»** μετά τον όρο **(3 ιδιότητες)**. Θα ήταν καλύτερα να επιτρεπόταν και ο ρηματικός τύπος «has» αλλά το σύστημα που χρησιμοποιήσαμε στα πειράματα δεν το επέτρεπε.

    Παράδειγμα : *«Amphibians can live both on land and in water»*

18. Ένα από τα ρήματα **«called», «known», «defined»** πριν τον όρο **(3 ιδιότητες)**

    Παράδειγμα : *«The giant wave known as tsunami»*

2.2.3.2 Αυτόματα επιλεγμένες ιδιότητες

Οι υπόλοιπες ιδιότητες είναι επίσης δυαδικές. Αφορούν και αυτές φράσεις, οι οποίες στην περίπτωση αυτή είναι ν-γράμματα (ακολουθίες) λέξεων που προηγούνται ή ακολουθούν αμέσως τον όρο-στόχο. Η διαφορά αυτών των ιδιοτήτων από τις προηγούμενες είναι πως οι φράσεις στις οποίες αντιστοιχούν επιλέγονται αυτόματα από το σύστημα. Η διαδικασία επιλογής επηρεάζεται πολύ από το περιεχόμενο των ερωτήσεων και μπορεί έτσι να προκύψουν διαφορετικές ιδιότητες, αν το σύστημα εκπαιδευθεί σε ερωτήσεις π.χ. ιατρικού ή γεωγραφικού περιεχομένου.

Το πλήθος των ιδιοτήτων αυτού του είδους είναι παράμετρος του συστήματος, η τιμή της οποίας προσδιορίζεται κατά την εκπαίδευση. Η επιλογή των φράσεων γίνεται από τα παράθυρα εκπαίδευσης, όπως στην εργασία της Μηλιαράκη [Mi03]:

- Δημιουργείται μια κενή λίστα φράσεων.
- Από όλα τα παράθυρα εκπαίδευσης (για όλους τους όρους-στόχους εκπαίδευσης) εξάγονται όλα τα ν-γράμματα που προηγούνται ή ακολουθούν τον όρο-στόχο. Τα ν-γράμματα που βρέθηκαν πριν τον όρο-στόχο θεωρούνται διαφορετικά από τα ν-γράμματα που βρέθηκαν μετά τον όρο-στόχο, ακόμα και αν αποτελούνται από τις ίδιες ακριβώς λέξεις. Για κάθε ν-

γραμμα που βρέθηκε πριν τον (οποιονδήποτε) όρο-στόχο, μετράμε και πόσες φορές βρέθηκε πριν τον όρο στόχο. Ομοίως για κάθε ν-γραμμα που βρέθηκε μετά τον όρο-στόχο.

- Ως αποτέλεσμα έχουμε μία λίστα με όλες τις φράσεις που έχουν εμφανιστεί αμέσως πριν ή μετά από οποιονδήποτε όρο-στόχο εκπαίδευσης, καθώς και πόσες φορές έχει εμφανιστεί η καθεμία πριν ή μετά από τον όρο-στόχο.

  Αυτή η λίστα είναι προφανώς πάρα πολύ μεγάλη. Για αυτόν το λόγο, φροντίζουμε να διαγράψουμε από αυτή κάθε φράση που εμφανίζεται λιγότερες φορές από κάποιο κατώφλι στα παράθυρα εκπαίδευσης.

- Μετά υπολογίζουμε την ακρίβεια (precision) κάθε φράσης (ν-γράμματος) που έχει απομείνει στην λίστα. Η ακρίβεια υπολογίζεται ως ο λόγος των παραθύρων όπου εμφανίζεται η φράση και είναι ορισμοί, δια τα συνολικά παράθυρα στα οποία εμφανίζεται η φράση. Η ακρίβεια μας δείχνει κατά πόσο η εμφάνιση της φράσης σηματοδοτεί με βεβαιότητα ότι το παράθυρο είναι ορισμός.

  Εδώ παρατηρούμε ότι αν μια φράση εμφανίζεται μόνο μια φορά σε ένα παράθυρο και αν αυτό το παράθυρο είναι και ορισμός, τότε η ακρίβεια αυτής της φράσης είναι 1. Αυτή όμως η φράση δεν δίνει σημαντική πληροφορία στο σύστημά μας, γιατί είναι πολύ σπάνια. Αυτός είναι ένας ακόμα λόγος που διαγράφουμε στο προηγούμενο βήμα τις φράσεις που εμφανίζονται πάρα πολύ λίγες φορές.

- Στο τέλος, ταξινομούμε τις φράσεις (ιδιότητες) κατά φθίνουσα ακρίβεια και επιλέγουμε τις κορυφαίες m, όπου m παράμετρος του συστήματος.

Η αξιολόγηση των ιδιοτήτων μπορεί να γίνει και βάσει άλλων μέτρων, πέραν της ακρίβειας. Στα πειράματα της εργασίας δοκιμάσαμε ακόμα την ανάκληση, που ορίζεται ως ο λόγος των παραθύρων όπου εμφανίζεται η φράση της ιδιότητας και είναι ορισμοί δια το συνολικό αριθμό παραθύρων που είναι ορισμοί.

2.2.4 Εκπαίδευση και ταξινόμηση

Όπως αναφέραμε και προηγουμένως, πρώτα δίνουμε στο σύστημα τις ερωτήσεις εκπαίδευσης. Έχοντας όλα τα παράθυρα που προέκυψαν από αυτές κωδικοποιημένα ως διανύσματα και έχοντας σημειώσει το καθένα από αυτά με την κατηγορία του (ορισμός ή μη) με έναν από τους τρόπους που περιγράψαμε παραπάνω, τα δίνουμε στην (κύρια) Μ.Δ.Υ. για να εκπαιδευθεί. Το σύστημα της παρούσας εργασίας χρησιμοποιεί την υλοποίηση LibSVM [ChLi01] των Μ.Δ.Υ., με πυρήνα RBF, η οποία επιστρέφει για κάθε απόφαση της Μ.Δ.Υ. και ένα βαθμό βεβαιότητας. Για την εξεύρεση των καλύτερων τιμών των παραμέτρων της Μ.Δ.Υ. χρησιμοποιείται η μέθοδος grid search που παρέχουν οι κατασκευαστές του LibSVM.

Μετά την εκπαίδευση, μπορούμε να δώσουμε στο σύστημα ένα σύνολο ερωτήσεων ορισμού που θέλουμε να απαντήσει. Το σύστημα θα κάνει την ίδια διαδικασία για κάθε ερώτηση, παράγοντας

πάλι διανύσματα τα οποία θα ταξινομηθούν ως ορισμοί ή μη ορισμοί από την εκπαιδευμένη πλέον Μ.Δ.Υ. Για κάθε ερώτηση, το σύστημα θα επιλέξει τα παράθυρα που τα αντίστοιχά τους διανύσματα θεωρήθηκαν ως τα πιο πιθανά να είναι ορισμοί και θα τα επιστρέψει στον χρήστη. Ο αριθμός των επιστρεφομένων παραθύρων ανά ερώτηση είναι επίσης παράμετρος του συστήματος.

2.3 Απαντήσεις αποτελούμενες από πολλαπλά αποσπάσματα

Όπως αναφέραμε στις προηγούμενες παραγράφους, τα συστήματα ερωταποκρίσεων μεμονωμένου αποσπάσματος προσπαθούν να απαντήσουν μια ερώτηση φυσικής γλώσσας του χρήστη επιστρέφοντας Ν σύντομα αποσπάσματα (συνήθως προτάσεις ή φράσεις, στην περίπτωσή μας παράθυρα) ως υποψήφιες απαντήσεις. Τα επιστρεφόμενα αποσπάσματα αποτελούν ουσιαστικά Ν ευκαιρίες του συστήματος να απαντήσει σωστά. Το σύστημα δεν κάνει καμία περαιτέρω επεξεργασία στα Ν αποσπάσματα, που μπορεί να παρουσιάζουν επικαλύψεις (π.χ. να επαναλαμβάνουν τις ίδιες πληροφορίες).

Υπάρχουν όμως ερωτήσεις που δεν μπορούν να απαντηθούν επαρκώς με ένα μοναδικό απόσπασμα. Ένα παράδειγμα στην περίπτωση των ερωτήσεων ορισμού είναι οι ασθένειες. Θα μπορούσαμε ίσως να συμφωνήσουμε ότι οι ασθένειες ορίζονται από τα συμπτώματά τους, αλλά πολλοί θεωρούν ίσης σημασίας τις επιπτώσεις τους και τις αιτίες πρόκλησής τους, πληροφορίες που μπορεί να περιλαμβάνονται σε διαφορετικά αποσπάσματα. Παρόμοια, στις ερωτήσεις που αφορούν πρόσωπα, εξίσου σημαντικά μπορούν να θεωρηθούν διαφορετικά στοιχεία της ζωής ενός προσώπου, όπως η αιτία θανάτου, το επάγγελμα κ.τ.λ., πληροφορίες που και πάλι ενδέχεται να περιλαμβάνονται σε διαφορετικά αποσπάσματα. Για παράδειγμα, παρατηρήστε τα παρακάτω παράθυρα που εξάγει από το διαδίκτυο το σύστημά μας για τον όρο-στόχο «ελονοσία» («malaria»).

Όρος-στόχος:

Malaria

Παράθυρα που εξήγαγε το σύστημα:

(…) malaria is a mosquito-borne disease caused by a parasite (…)

(…) people with malaria often experience fever, chills, and flu-like illness. left untreated, they may develop severe complications and die.(…)

(…) each year 350-500 million cases of malaria occur worldwide, and over one million people die, most of them young (…)

(…) people who get malaria are typically very sick with high fevers, shaking chills, and flu-like illness. (…)

(…) there are four different types of malaria caused by four related parasites. (…)

Ο χρήστης πιθανότατα ενδιαφέρεται για όλες τις πληροφορίες που εμφανίζονται στα παραπάνω παράθυρα. Θα ήταν προτιμότερο, επομένως, να επιστρέφεται ως απάντηση ένας συνδυασμός πολλών παραθύρων, δηλαδή να επιτρέψουμε η κάθε απάντηση του συστήματος να αποτελείται από πολλά παράθυρα, που να συνθέτουν έναν κατά το δυνατόν πλήρη ορισμό του όρου-στόχου. Στην περίπτωση αυτή, μιλάμε για συστήματα ερωταποκρίσεων πολλαπλών αποσπασμάτων (multi-snippet question answering systems).

2.3.1 Αλλαγές στο σύστημα της εργασίας

Στα στάδια που εξετάσαμε στις προηγούμενες παραγράφους, το σύστημά μας εξάγει από το διαδίκτυο έγγραφα και στη συνέχεια από αυτά παράθυρα που περιέχουν τον όρο-στόχο. Έπειτα, αυτά τα παράθυρα αξιολογούνται και σε κάθε ένα αντιστοιχείται ένας βαθμός (score) που δείχνει πόσο πιθανό είναι το παράθυρο να περιέχει πληροφορία που ορίζει τον όρο-στόχο. Τα παράθυρα με τους Ν υψηλότερους βαθμούς επιλέγονται ως υποψήφιες απαντήσεις.

Στην περίπτωση που θέλουμε τα επιστρεφόμενα παράθυρα να συνθέτουν μια ολοκληρωμένη, πιο εκτεταμένη απάντηση του συστήματος και να μην αποτελούν απλά Ν ευκαιρίες του συστήματος να απαντήσει σωστά, ακολουθεί ένα νέο στάδιο: αν θεωρήσουμε ότι η ολοκληρωμένη απάντηση του συστήματος θέλουμε να έχει μέγεθος Κ παραθύρων, τότε στόχος του νέου σταδίου είναι να επιλεγούν Κ από τις υποψήφιες Ν απαντήσεις (παράθυρα), ώστε οι πληροφορίες που μεταφέρουν τα Κ παράθυρα να ορίζουν τον όρο-στόχο, χωρίς να επαναλαμβάνονται πληροφορίες και χωρίς να περιλαμβάνονται άσχετες πληροφορίες. Στις επόμενες ενότητες θα παρουσιάσουμε διάφορες μεθόδους που μπορούν να χρησιμοποιηθούν σε αυτό το νέο στάδιο.

2.3.2 Μέτρα σύγκρισης

Η πιο συνηθισμένη μέθοδος επιλογής των Κ παραθύρων (γενικότερα υποψηφίων απαντήσεων) στη βιβλιογραφία είναι η σύγκριση των Ν παραθύρων μεταξύ τους ως προς το περιεχόμενό τους. Αν δύο παράθυρα βρεθούν, χρησιμοποιώντας κάποιο μέτρο, πολύ όμοια, τότε μπορούμε να θεωρήσουμε ότι περιέχουν την ίδια πληροφορία, οπότε μόνο ένα από αυτά χρειάζεται να συμπεριληφθεί στην τελική απάντηση. Το παράθυρο που επιλέγεται από τα δύο είναι αυτό που έχει το μεγαλύτερο βαθμό (score), όπως αυτός υπολογίστηκε στα προηγούμενα βήματα.

Η διαδικασία ξεκινάει με το παράθυρο με το μεγαλύτερο βαθμό να προστίθεται στο σύνολο Σ των παραθύρων που θα περιληφθούν στην τελική παράγραφο. Σε κάθε βήμα εξετάζουμε το αμέσως επόμενο παράθυρο στη κατάταξη (ως προς το βαθμό) και το συγκρίνουμε με όσα έχουν ήδη προστεθεί στο σύνολο Σ. Αν το παράθυρο δεν είναι όμοιο με κανένα από τα παράθυρα του Σ, προστίθεται και αυτό στο Σ· διαφορετικά το αγνοούμε και συνεχίζουμε εξετάζοντας το επόμενο στη κατάταξη. Η διαδικασία επαναλαμβάνεται μέχρι το Σ να περιέχει Κ παράθυρα ή να εξαντλήσουμε τα Ν παράθυρα.

Θα δούμε τώρα μερικά απλά μέτρα σύγκρισης των παραθύρων. Το πρώτο είναι η απόσταση συνημίτονου μεταξύ των παραθύρων, όπως ορίστηκε στην παράγραφο 2.2.2.3. Όπως και πριν, η απόσταση υπολογίζεται μεταξύ ενός παραθύρου και κάθε παραθύρου που ανήκει ήδη στο σύνολο Σ. Αν η απόσταση βρεθεί μεγαλύτερη από κάποιο προκαθορισμένο κατώφλι t, τότε το παράθυρο αγνοείται και προχωράμε στο επόμενο.

Μια παραλλαγή της απόστασης συνημίτονου είναι η σταθμισμένη απόσταση συνημίτονου με βάρη TF-IDF. Η τιμή TF (term frequency) δηλώνει τη συχνότητα εμφανίσεως μιας λέξης σε ένα κείμενο ή στην περίπτωση μας στο σύνολο των παραθύρων που εξήχθησαν για τον όρο στόχο. Οι τιμές IDF υπολογίζονται όπως δείξαμε σε προηγούμενη παράγραφο.

$$tfidf_i = tf_i \cdot idf_i = \frac{N_i}{N} \cdot idf_i$$

Όπου:

N_i: το σύνολο των εμφανίσεων της λέξης i στα παράθυρα που εξήχθησαν για τον όρο-στόχο

N: το πλήθος των λέξεων στα παράθυρα που εξήχθησαν για τον όρο-στόχο

Στην συνέχεια, τα παράθυρα προς σύγκριση μετατρέπονται σε διανύσματα, όλα τόσων διαστάσεων όσες είναι συνολικά οι λέξεις που εμφανίζονται σε αυτά. Η διαφορά από τα αντίστοιχα διανύσματα της απλής απόστασης συνημιτόνου είναι πως οι ιδιότητες των διανυσμάτων δεν είναι τώρα δυαδικές, αλλά περιέχουν τις τιμές tf-idf για κάθε λέξη. Αφού κατασκευαστούν τα διανύσματα, η απόσταση συνημίτονου υπολογίζεται με τον ίδιο τύπο:

$$\text{score} = \frac{\text{vector}_A * \text{vector}_B}{|\text{vector}_A| + |\text{vector}_B|}$$

Όπου:

vector: τα διανύσματα των παραθύρων

* το εσωτερικό γινόμενο,

|vector|: το μέτρο του διανύσματος.

2.4 Περιγραφή άλλων συστημάτων

Παρακάτω περιγράφουμε συνοπτικά μερικά από τα συστήματα άλλων ερευνητών που έχουν προταθεί στα πλαίσια των διαγωνισμών TREC και άλλων συνεδρίων. Στα πειράματα των επομένων κεφαλαίων θα συγκρίνουμε τις επιδόσεις αυτών των συστημάτων με τις επιδόσεις του δικού μας. Τα

συγκεκριμένα συστήματα επελέγησαν κυρίως επειδή είχαν επιτύχει πολύ υψηλές επιδόσεις στα συνέδρια όπου πήραν μέρος. Εξαιρέσαμε συστήματα που χρησιμοποιούν γλωσσάρια και εγκυκλοπαίδειες κατά τη χρήση τους (όχι μόνο κατά την εκπαίδευσή τους), μια που ο σκοπός μας είναι να βρίσκουμε κατά τη χρήση του συστήματος ορισμούς όρων που δεν περιλαμβάνονται σε εγκυκλοπαίδειες.

2.4.1 Cui κ.ά. - Πανεπιστήμιο Σιγκαπούρης

Το σύστημα των Cui κ.ά. [Cu06] λειτουργεί σε γενικές γραμμές ως εξής. Αρχικά, για κάθε όρο-στόχο, αναζητούνται σε μια συλλογή κειμένων οι προτάσεις που τον περιέχουν. Από τις λέξεις των προτάσεων αυτών δημιουργείται ένα κεντροειδές, όπως στην ενότητα 2.2.2.3. Κατόπιν μετράται, με το μέτρο της ομοιότητας συνημίτονου, η απόσταση κάθε προτάσεως από το κεντροειδές και επιλέγονται ως πιθανές προτάσεις ορισμού οι προτάσεις που βρίσκονται εγγύτερα στο κεντροειδές.

Οι προτάσεις που επελέγησαν περνούν στη συνέχεια από ένα σύστημα αναγνώρισης μερών του λόγου (part-of-speech tagger) και ένα σύστημα ρηχής συντακτικής ανάλυσης (chunker). Οι λέξεις του όρου-στόχου αντικαθίστανται στις προτάσεις από την ψευδο-λέξη «<TARGET>». Οι λέξεις που συμμετέχουν στο κεντροειδές αντικαθίστανται στις προτάσεις από τα μέρη του λόγου στα οποία ανήκουν (POS tags). Οι ονοματικές φράσεις (noun phrases) αντικαθίστανται από την ψευδο-λέξη «NP», οι τύποι του ρήματος «to be» από την ψευδο-λέξη «BE$», τα άρθρα «a», «an», «the» από την ψευδο-λέξη «DT$», οι αριθμητικές τιμές από την «CD$», τα επίθετα και επιρρήματα αφαιρούνται, ενώ οι υπόλοιπες λέξεις παραμένουν ως έχουν. Στη συνέχεια η πρόταση χωρίζεται σε μια δεξιά και αριστερή ακολουθία περί τη θέση του όρου-στόχου. Τέλος κλαδεύουμε τις δύο ακολουθίες, ώστε να έχουν μέγιστο πλήθος στοιχείων L η κάθε μία, όπου L παράμετρος του συστήματος.

Προκύπτουν έτσι παράθυρα του όρου-στόχου, μήκους το πολύ 2L+1 λέξεων, τα οποία αξιολογούνται κατόπιν ως προς την πιθανότητά τους να αντιστοιχούν σε ορισμούς. Οι Cui κ.ά. προτείνουν δύο τρόπους αξιολόγησης, έναν που χρησιμοποιεί ένα μοντέλο διγραμμάτων (Bigram Model, BM) και έναν που χρησιμοποιεί ένα πιο περίπλοκο μοντέλο, το οποίο ονομάζουν Profile Hidden Markov Model (PHMM). Ο δεύτερος τρόπος (PHMM) οδήγησε σε καλύτερα αποτελέσματα στα πιο πρόσφατα πειράματα των Cui κ.ά. [Cu06], αλλά η διαφορά από τα αποτελέσματα του πρώτου τρόπου (BM) δεν ήταν στατιστικά σημαντική, ενώ σε παλαιότερη εργασία τους [CuKa05], όπου χρησιμοποιούνταν λιγότερα πειραματικά δεδομένα, ο πρώτος τρόπος (BM) υπερτερούσε. Ως εκ τούτου, στην παρούσα εργασία ασχολούμαστε μόνο με το BM, αφού είναι απλούστερο και οδηγεί σε περίπου τα ίδια αποτελέσματα.

Στο BM, για κάθε ένα παράθυρο αξιολογούνται ξεχωριστά η δεξιά και αριστερή του ακολουθία, κάθε μία από τις οποίες αποτελείται από L το πολύ λέξεις. Συμβολίζουμε, όπως οι Cui κ.ά., με t_1, …, t_L τις συγκεκριμένες λέξεις κάθε μίας από τις δύο ακολουθίες σε ένα συγκεκριμένο παράθυρο, ενώ με S_1, …, S_L συμβολίζουμε τις θέσεις (slots) των δύο ακολουθιών εν γένει.

Αρχικά, εκτιμώνται οι ακόλουθες πιθανότητες, ξεχωριστά για αριστερές και δεξιές ακολουθίες. Στην περίπτωση της $P(t_i|S)$, οι όροι δ και δN χρησιμοποιούνται για εξομάλυνση, επειδή διαφορετικά οι εκτιμήσεις είναι συχνά μηδενικές. Στην περίπτωση της $P(t_i|t_i\text{-}1)$, χρησιμοποιείται απλά εκτίμηση μέγιστης πιθανοφάνειας:

$$P\left(t_i \middle| S_i\right) = \frac{\left|t_i\left(S_i\right)\right| + \delta}{\sum_t \left|t\left(S_i\right)\right| + \delta \cdot N}$$

$$P\left(t_i \middle| t_{i-1}\right) = \frac{\left|t_i\left(S_i\right) t_{i-1}\left(S_{i-1}\right)\right|}{\left|t_i\left(S_i\right)\right|}$$

Όπου:

$P_{ML}(t_i|S_i)$: η πιθανότητα η λέξη t_i να εμφανίζεται στη θέση S_i μιας αριστερής ή δεξιάς, ανάλογα με την περίπτωση, ακολουθίας παραθύρου ορισμού

$P_{ML}(t_i|t_{i-1})$: η πιθανότητα η λέξη t_i να ακολουθεί την t_{i-1} σε μια αριστερή ή δεξιά, ανάλογα με την περίπτωση, ακολουθία παραθύρου ορισμού

$|t(S_i)|$: το πλήθος των εμφανίσεων της λέξης t στη θέση S_i των αριστερών ή δεξιών, ανάλογα με την περίπτωση, ακολουθιών των παραθύρων ορισμού εκπαίδευσης

N: το πλήθος των διαφορετικών λέξεων που εμφανίζονται στα παράθυρα ορισμού εκπαίδευσης

δ: μια σταθερά, με τιμή 2 στα πειράματα των Cui κ.ά.

$|t_i(S_i)\ t_{i-1}(S_{i-1})|$: το πλήθος των συν-εμφανίσεων των λέξεων t_i και t_{i-1} στις θέσεις S_i και S_{i-1}, αντίστοιχα, των αριστερών ή δεξιών, ανάλογα με την περίπτωση, ακολουθιών των παραθύρων εκπαίδευσης

Παρατηρούμε ότι η μέθοδος των Cui κ.ά. χρησιμοποιεί μόνο θετικά παραδείγματα (παραδείγματα ορισμών, όχι μη-ορισμών). Οι Cui κ.ά. αναφέρουν, επίσης, ότι κατά την εκτίμηση των $P(t_i|S_i)$, αν η t_i είναι ψευδο-λέξη, τότε λαμβάνουν υπόψη τους μόνο τις εμφανίσεις ψευδο-λέξεων στα παραδείγματα εκπαίδευσης. Αντίστοιχα, αν η t_i είναι πραγματική λέξη, τότε λαμβάνουν υπόψη τους μόνο τις εμφανίσεις πραγματικών λέξεων στα παραδείγματα εκπαίδευσης. Αυτό το κάνουν επειδή οι ψευδο-λέξεις είναι πολύ πιο συχνές, σε βαθμό που οι εκτιμήσεις των $P(t_i|S_i)$ για πραγματικές λέξεις θα γίνονταν περίπου μηδενικές χωρίς τον παραπάνω διαχωρισμό.

Στη συνέχεια εκτιμάται η πιθανότητα της κάθε ακολουθίας ως εξής:

$$\mathrm{P}(t_1 \ldots t_L) = \mathrm{P}(t_1 \mid S_1) \prod_{i=2}^{L} \left(\lambda \mathrm{P}(t_i \mid t_{i-1}) + (1-\lambda) \mathrm{P}(t_i \mid S_i) \right)$$

Όπου:

λ, το βάρος που συνδυάζει τις $P(t_i|t_{i-1})$ και $P(t_i|S_i)$.

Επειδή κάποιες ακολουθίες είναι μικρότερες από L (π.χ. προήλθαν από μικρότερες προτάσεις ή λόγω της αντικατάστασης ονοματικών φράσεων με NP), αντί της πιθανότητας του παραπάνω τύπου χρησιμοποιείται η αντίστοιχη πιθανοφάνεια, κανονικοποιημένη ως προς το μήκος l του παραθύρου που αξιολογείται.

$$P_{norm}(t_1 \ldots t_l) = \frac{1}{l} \left(\log \mathrm{P}(t_1 \mid S_1) + \sum_{i=2}^{l} \log \left(\lambda \mathrm{P}(t_i \mid t_{i-1}) + (1-\lambda) \mathrm{P}(t_i \mid S_i) \right) \right)$$

Το βάρος λ στους παραπάνω τύπους εκτιμάται με τον αλγόριθμο Μεγιστοποίησης Αναμονής (EM – Expectation Maximization), ο οποίος επιχειρεί να μεγιστοποιήσει την πιθανοφάνεια όλων των παραδειγμάτων εκπαίδευσης.

$$\lambda = \arg\max_{\lambda'} \sum_{j=1}^{|INS|} \mathrm{P}_{norm}\left(t_1^{(j)} \ldots t_{l(j)}^{(j)} \mid \lambda' \right)$$

Όπου:

INS: τα παράθυρα εκπαίδευσης

|INS|: το πλήθος των παραδειγμάτων εκπαίδευσης.

Τα βήματα του αλγορίθμου ΕΜ είναι τα εξής:

1. Αρχικοποιούμε το λ σε μια τυχαία τιμή από το 0 ως το 1.
2. Επανεκτιμούμε το λ με βάση τον τύπο:

$$\lambda \leftarrow \frac{1}{|INS|} \cdot \sum_{j=1}^{|INS|} \frac{1}{l_{(j)} - 1} \sum_{i=2}^{l_{(j)}} \frac{\lambda \mathrm{P}\left(t_i^{(j)} \mid t_{i-1}^{(j)} \right)}{\lambda \mathrm{P}\left(t_i^{(j)} \mid t_{i-1}^{(j)} \right) + (1-\lambda) \mathrm{P}\left(t_i^{(j)} \mid S_i^{(j)} \right)}$$

3. Επαναλαμβάνουμε το βήμα 2 μέχρι να συγκλίνει το λ.

Στα πειράματα των Cui κ.α. το λ συνέκλινε στην τιμή 0,3.

Τέλος, αφού έχουμε υπολογίσει τις πιθανότητες για την δεξιά και αριστερή ακολουθία, συνδυάζουμε γραμμικά τα αποτελέσματα:

$$\mathrm{P} = (1-a)\,\mathrm{P}(left) + a\,\mathrm{P}(right)$$

Όπου:

P: η πιθανότητα να είναι ορισμός το παράθυρο

P(left): η πιθανότητα της αριστερής ακολουθίας

Pr(right): η πιθανότητα της δεξιάς ακολουθίας

α: το βάρος που συνδυάζει τις πιθανότητες

Το α μπορεί και αυτό να εκτιμηθεί με τον αλγόριθμο EM. Οι Cui κ.α. χρησιμοποιούσαν στα πειράματα τους τη τιμή 0,7.

2.4.2 Xu κ.ά. - BBN

Οι Xu κ.ά. [XuWe04] χρησιμοποιούν τη μηχανή εξαγωγής πληροφοριών Serif [RaBo01], η οποία, μεταξύ άλλων λειτουργιών, εκτελεί συντακτική ανάλυση (parsing) και επιλύει αναφορικές εκφράσεις (anaphora resolution, π.χ. αντωνυμιών). Σε κάθε ερώτηση, οι Xu κ.ά. δίνουν τον όρο-στόχο της ερώτησης σε μια μηχανή ανάκτησης πληροφοριών και συλλέγουν τα 1.000 σχετικότερα έγγραφα. Τροφοδοτούν κατόπιν με τα έγγραφα αυτά τη μηχανή Serif και αξιοποιώντας τα αποτελέσματά της (συντακτικά δέντρα, επίλυση αναφορικών εκφράσεων) εξάγουν από τα έγγραφα προτάσεις που περιέχουν τον όρο-στόχο ή αναφορές σε αυτόν. Στις προτάσεις αυτές, εντοπίζουν πέντε κατηγορίες φράσεων, τις οποίες ονομάζουν «χαρακτηριστικά» (features):

- Παραθέσεις (appositives, π.χ. «George Bush, the US president»).
- Ονοματικές φράσεις κατηγορημάτων (π.χ. «George Bush is the US president»).
- Φράσεις που αντιστοιχούν σε συγκεκριμένες θέσεις (slots) προτύπων (patterns) τα οποία ταιριάζουν με τα συντακτικά δέντρα των προτάσεων. Για παράδειγμα, στην περίπτωση της πρότασης «[...] suffering from symptoms of depression and exhaustion , an illness similar to **Gulf War syndrome** […]» και του προτύπου «(a|an) NP similar to QTERM», όπου QTERM είναι ο όρος-στόχος, εξάγεται η υπογραμμισμένη ονοματική φράση. Ο όρος-στόχος είναι σημειωμένος με έντονα γράμματα. Οι Xu κ.ά. αναφέρουν ότι χρησιμοποιούν περισσότερα από 40 πρότυπα, τα οποία υποθέτουμε ότι κατασκεύασαν χειρωνακτικά.

- Φράσεις που αντιστοιχούν σε λογικά κατηγορήματα (π.χ. went(Smith, Spain)) τα οποία εντοπίζει η μηχανή Serif στις προτάσεις. Οι Xu κ.ά. αποδίδουν ιδιαίτερη βαρύτητα σε φράσεις των οποίων τα κατηγορήματα ταιριάζουν με κάποιο από τα περίπου 100 σχεδιότυπα (templates) κατηγορημάτων που έχουν ορίσει οι ίδιοι. Οι φράσεις αυτές χαρακτηρίζονται «ειδικές» («special»), ενώ οι υπόλοιπες φράσεις που αντιστοιχούν σε κατηγορήματα χαρακτηρίζονται «γενικές» («generic»).
- Φράσεις που αντιστοιχούν στις 24 δυαδικές σχέσεις (π.χ. υπάλληλος-του, γονέας-του) του ερευνητικού προγράμματος ACE [LDC02], τις οποίες υποστηρίζει η μηχανή Serif.

Χρησιμοποιούνται, ακόμη, ως χαρακτηριστικά τελευταίας επιλογής και οι ίδιες οι προτάσεις, όπως εξηγείται παρακάτω.

Για κάθε ερώτηση, κατασκευάζεται επίσης ένα κεντροειδές. Το κεντροειδές περιλαμβάνει τις λέξεις που εμφανίζονται στους ορισμούς του όρου-στόχου σε ηλεκτρονικές εγκυκλοπαίδειες, γλωσσάρια κλπ. Αν δεν βρεθούν ορισμοί του όρου-στόχου σε τέτοιες πηγές, τότε: (α) αν ο όρος-στόχος είναι όνομα προσώπου, χρησιμοποιείται το κεντροειδές όλων των ορισμών μιας συλλογής βιογραφιών, (β) διαφορετικά χρησιμοποιείται ένα κεντροειδές που περιλαμβάνει όλα τα χαρακτηριστικά που εξήχθησαν από όλες τις προτάσεις της συγκεκριμένης ερώτησης.

Σε κάθε ερώτηση, υποψήφιες απαντήσεις είναι τα χαρακτηριστικά που εξήχθησαν από τις προτάσεις της. Μεταξύ αυτών, επιλέγονται πρώτα τα 10 χαρακτηριστικά που έχουν το μεγαλύτερο βαθμό ομοιότητας με το κεντροειδές. Τα υπόλοιπα χαρακτηριστικά κατατάσσονται στις εξής κατηγορίες: (i) φράσεις από παραθέσεις και κατηγορήματα, (ii) φράσεις από θέσεις προτύπων, (iii) «ειδικές» φράσεις από κατηγορήματα, (iv) φράσεις από σχέσεις, (v) «γενικές» φράσεις από κατηγορήματα και ολόκληρες προτάσεις. Από κάθε κατηγορία, επιλέγονται τα χαρακτηριστικά που παρουσιάζουν το μεγαλύτερο βαθμό ομοιότητας με το κεντροειδές και προστίθενται στα αρχικά 10. Τα επιλεγμένα χαρακτηριστικά επιστρέφονται στο χρήστη.

Οι Xu κ.ά. αναφέρουν ότι κατά τον υπολογισμό της ομοιότητας κάθε χαρακτηριστικού με το κεντροειδές, χρησιμοποιούν τις τιμές TF-IDF των χαρακτηριστικών, χωρίς όμως να παρέχουν περισσότερες πληροφορίες για το πώς ακριβώς προκύπτουν αυτές οι τιμές στην περίπτωση της μεθόδου τους. Δεν εξηγούν, επίσης, ούτε ποιο ακριβώς μέτρο ομοιότητας χρησιμοποιούν με τις τιμές TF-IDF, παραπέμποντας στην εργασία [AlCa20].

Σημειώνουμε, τέλος, ότι οι Xu κ.ά. χρησιμοποιούν πρόσθετα κριτήρια εντοπισμού ισοδύναμων χαρακτηριστικών (φράσεων), ώστε να μην περιληφθούν οι ίδιες πληροφορίες πολλές φορές στη συνολική απάντηση που επιστρέφεται στο χρήστη.

- Στην περίπτωση χαρακτηριστικών που αντιστοιχούν σε λογικά κατηγορήματα, δύο χαρακτηριστικά θεωρούνται ισοδύναμα αν, χονδρικά, έχουν το ίδιο ρήμα και τα ίδια κύρια ουσιαστικά (head nouns) στο υποκείμενο και το αντικείμενό τους.

- Στην περίπτωση χαρακτηριστικών που προέρχονται από θέσεις προτύπων, δύο χαρακτηριστικά θεωρούνται ισοδύναμα αν προέρχονται από το ίδιο πρότυπο (υποθέτουμε και την ίδια θέση).
- Σε όλες τις άλλες περιπτώσεις χαρακτηριστικών, αν περισσότερες από 70% των λέξεων του χαρακτηριστικού περιλαμβάνονται σε άλλα χαρακτηριστικά που έχουν ήδη επιλεγεί, τότε το χαρακτηριστικό αγνοείται.

Τελικά επιστρέφεται ένας συγκεκριμένος συνολικός αριθμός χαρακτηριστικών ή χαρακτήρων (π.χ. 10 χαρακτηριστικά ή μέχρι 4000 χαρακτήρες).

2.4.3 Blair-Goldensohn κ.ά. – Columbia

Οι Blair-Goldensohn κ.ά χρησιμοποιούν ένα συνδυασμό τεχνικών μηχανικής μάθησης και συντακτικών/λεκτικών προτύπων στο σύστημα DefScriber [BGMc03] [BGMc04] . Για κάθε όρο-στόχο ανακτάται ένας αριθμός ιστοσελίδων μέσω μιας μηχανής αναζήτησης. Οι ιστοσελίδες στη συνέχεια αναλύονται και εντοπίζονται σε αυτές «κατηγορήματα ορισμού» (ουσιαστικά προτάσεις) τριών ειδών: κατηγορήματα «γένους» (genus), κατηγορήματα «είδους» (species) και γενικά κατηγορήματα ορισμού. Τα κατηγορήματα γένους αντιστοιχούν σε σχέσεις «is-a» μεταξύ του όρου-στόχου και κάποιας κατηγορίας ή συνόλου (π.χ. «The Hajj is a type of ritual»). Τα κατηγορήματα είδους περιέχουν πληροφορίες διαφορετικές ή επιπρόσθετες του γένους (π.χ. «The annual Hajj begins in the twelfth month of the Islamic year»). Οι Blair-Goldensohn κ.ά. δεν ορίζουν με επαρκή σαφήνεια την έννοια των κατηγορημάτων είδους. Τέλος, τα γενικά κατηγορήματα είναι υπερ-σύνολο των κατηγορημάτων γένους και είδους. Αντιστοιχούν σε οποιαδήποτε πληροφορία μπορεί να περιληφθεί σε έναν εκτενή ορισμό του όρου-στόχου.

Κατά την ανάλυση των ιστοσελίδων, σε πρώτο στάδιο εντοπίζονται τα γενικά κατηγορήματα, ουσιαστικά οι προτάσεις που μπορούν εν γένει να περιληφθούν σε ορισμούς. Χρησιμοποιείται μηχανική μάθηση και πιο συγκεκριμένα το εργαλείο Ripper [Co95], με το οποίο οι Blair-Goldensohn κ.ά. κατασκευάζουν ένα δέντρο-απόφασης. Κάθε πρόταση παριστάνεται ως ένα διάνυσμα χαρακτηριστικών (διάνυσμα τιμών ιδιοτήτων) και το παραγόμενο δέντρο κατατάσσει τις προτάσεις (στην πραγματικότητα τα διανύσματά τους) ως γενικά κατηγορήματα ή όχι. Οι ιδιότητες επιλέγονται μετά από μελέτη εγγράφων εκπαίδευσης, στα οποία έχουν σημειωθεί χειρωνακτικά τα «κατηγορήματα». Για παράδειγμα, στα πειράματά τους οι Blair-Goldensohn κ.ά. χρησιμοποίησαν ιδιότητες όπως η συγκέντρωση του όρου-στόχου σε μια πρόταση (συχνότητα του όρου-στόχου στη πρόταση και στις γύρω προτάσεις), η σχετική και απόλυτη θέση της πρότασης σε ένα κείμενο, η εμφάνιση σημείων στίξης κ.ά. Βασικός σκοπός αυτού του σταδίου είναι να επικεντρωθεί η περαιτέρω αναζήτηση σε φράσεις που έχουν δυνητικώς σχέση με τον ορισμό του όρου-στόχου.

Στη συνέχεια, οι Blair-Goldensohn κ.ά. αναζητούν κατηγορήματα γένους, είδους και γένους-είδους (κατηγορήματα που περιέχουν ταυτόχρονα πληροφορίες γένους και είδους) ανάμεσα στα

γενικά κατηγορήματα του προηγούμενου σταδίου. Τα κατηγορήματα αυτά εντοπίζονται χρησιμοποιώντας 18 χειρωνακτικά κατασκευασμένα πρότυπα (patterns). Τα πρότυπα έχουν τη μορφή μερικώς καθορισμένων συντακτικών δέντρων στα οποία λέξεις ή φράσεις έχουν αντικατασταθεί με γενικότερες κατηγορίες (π.χ. ρήματα όπως «be» και «exemplify» αντικαθίστανται με «FormativeVb»).

Κατόπιν, κατασκευάζεται ένα κεντροειδές από τις λέξεις (αφού πρώτα αφαιρεθούν οι καταλήξεις τους) όλων των γενικών κατηγορημάτων που έχουν εντοπιστεί, τα γενικά κατηγορήματα ταξινομούνται κατά φθίνουσα ομοιότητα με το κεντροειδές και επιλέγονται τα κορυφαία (διαφορετικά) L κατηγορήματα. Η ομοιότητα των κατηγορημάτων με το κεντροειδές υπολογίζεται χρησιμοποιώντας τη σταθμισμένη απόσταση συνημίτονου με βάρη IDF (IDF-weighted cosine distance).

Έπειτα ομαδοποιούνται τα επιλεγέντα κατηγορήματα χρησιμοποιώντας ένα μη ιεραρχικό αλγόριθμο σειριακής ομαδοποίησης (clustering), για τον οποίο οι Blair-Goldensohn κ.ά. δεν παρέχουν περισσότερες πληροφορίες. Ως μέτρο απόστασης κατά την ομαδοποίηση χρησιμοποιείται πάλι η σταθμισμένη απόσταση συνημίτονου με βάρη IDF, με τη διαφορά ότι χρησιμοποιείται τώρα ένας συνδυασμός των τιμών IDF που προκύπτουν από μια γενική συλλογή εγγράφων (global IDF) και των τιμών IDF που προκύπτουν από τα γενικά κατηγορήματα του συγκεκριμένου όρου-στόχου (local IDF). Αντιπρόσωπος κάθε ομάδας (cluster) θεωρείται το κατηγόρημα-μέλος της ομάδας το οποίο έχει τη μεγαλύτερη ομοιότητα με το κεντροειδές.

Η τελική απάντηση (συνολικός ορισμός του όρου-στόχου) σχηματίζεται ως εξής. Τοποθετείται πρώτο στην απάντηση το κατηγόρημα γένους-είδους που έχει τη μεγαλύτερη ομοιότητα με το κεντροειδές. Οι υπόλοιπες προτάσεις της απάντησης επιλέγονται μεταξύ των αντιπροσώπων των ομάδων (clusters) σε διαδοχικά βήματα. Σε κάθε βήμα, επιλέγεται ο αντιπρόσωπος που έχει τη μικρότερη συνολική απόσταση από το κεντροειδές και από τον αντιπρόσωπο της ομάδας που επιλέχθηκε στο αμέσως προηγούμενο βήμα (και οι δύο αποστάσεις έχουν το ίδιο βάρος και υπολογίζονται με τον ίδιο μέτρο). Η διαδικασία επαναλαμβάνεται μέχρι να φτάσουμε στο επιθυμητό μέγεθος απάντησης.

2.4.4 Chu-Carroll κ.ά. - IBM

Το σύστημα PIQUANT της IBM [PrCh03] [ChCz04] [ChCz05] συνδυάζει συστατικά (components) διαφορετικών προσεγγίσεων, κάθε ένα από τα οποία εξειδικεύεται σε διαφορετικές κατηγορίες ερωτήσεων. Το σύστημα αναλύει κάθε ερώτηση που του δίνεται και τροφοδοτεί με αυτήν όλα τα συστατικά. Στη διάρκεια της ανάλυσης, επιλύονται αναφορικές εκφράσεις, συμπληρώνονται ελλειπτικές εκφράσεις και γενικά η ερώτηση μετασχηματίζεται σε αυτοτελή ερώτηση. Κάθε συστατικό επιστρέφει υποψήφιες απαντήσεις μαζί με ένα βαθμό βεβαιότητας (score) για κάθε μία. Οι βαθμοί βεβαιότητας των απαντήσεων από κάθε συστατικό πολλαπλασιάζονται με ένα μοναδικό για κάθε συστατικό βάρος, που αντιστοιχεί στην απόδοση αυτού κατά την διάρκεια προηγούμενων

πειραμάτων. Τέλος, επιστρέφονται οι απαντήσεις που έχουν το μεγαλύτερο συνολικό βαθμό βεβαιότητας.

Το πιο ενδιαφέρον και πιο σχετικό με ερωτήσεις ορισμών συστατικό λειτουργεί σε τέσσερα στάδια. Πρώτον, εξάγονται από τη συλλογή κειμένων κείμενα σχετικά με τον όρο-στόχο χρησιμοποιώντας τη μηχανή αναζήτησης JuruXML [CaMa03] και από κάθε κείμενο που εντοπίζει η JuruXML εξάγονται αποσπάσματα («passages») μήκους μίας προτάσεως το καθένα. (Οι Chu-Carroll κ.ά. δεν διευκρινίζουν πώς ακριβώς εξάγονται τα αποσπάσματα. Επίσης, στην πιο πρόσφατη εργασία τους αναφέρονται σε αποσπάσματα μήκους 1-3 προτάσεων το καθένα.) Δεύτερον, εντοπίζονται τα ουσιαστικά των αποσπασμάτων και υπολογίζονται οι συχνότητές τους. (Οι Chu-Carroll κ.ά. δεν διευκρινίζουν αν οι συχνότητες μετρούν σε πόσα αποσπάσματα εμφανίζεται το κάθε ουσιαστικό ή πόσες φορές συνολικά εμφανίζεται κάθε ουσιαστικό στο σύνολο των αποσπασμάτων.) Τα ουσιαστικά ταξινομούνται βάσει της διαφοράς μεταξύ της συχνότητας του καθενός και της αναμενόμενης συχνότητάς του, που προκύπτει από την τιμή IDF του ουσιαστικού. (Οι Chu-Carroll κ.ά. δεν διευκρινίζουν πώς ακριβώς υπολογίζουν τις τιμές IDF, ούτε πώς προκύπτουν οι αναμενόμενες συχνότητες από τις τιμές IDF.) Τα ουσιαστικά των οποίων η διαφορά συχνοτήτων υπερβαίνει ένα προκαθορισμένο κατώφλι θεωρούνται «υποψήφιες έννοιες» («candidate concepts»). Τρίτον, εντοπίζονται όλα τα αποσπάσματα, μεταξύ εκείνων που επέστρεψε η JuruXML, τα οποία περιέχουν την υποψήφια έννοια με τη μεγαλύτερη διαφορά συχνοτήτων. Από αυτά επιλέγεται το απόσπασμα που περιέχει το μεγαλύτερο αριθμό υποψηφίων εννοιών και επιστρέφεται. Όλες οι υποψήφιες έννοιες που περιέχονται σε αυτό το απόσπασμα διαγράφονται από τη λίστα των υποψηφίων εννοιών και η διαδικασία επαναλαμβάνεται μέχρι να αδειάσει η λίστα ή να φτάσουμε σε ένα προκαθορισμένο μέγιστο μέγεθος απάντησης. Στην πιο πρόσφατη εργασία τους, οι Chu-Carroll κ.ά. αναφέρουν ότι δεν χρησιμοποιούν πλέον μόνο ουσιαστικά, αλλά όλες τις «έννοιες» («concepts»), επιβραβεύοντας μάλιστα έννοιες που συνδέονται συντακτικά με τον όρο-στόχο στα αποσπάσματα. Δεν παρέχουν, όμως, περισσότερες πληροφορίες επ' αυτού.

Ένα άλλο συστατικό προσπαθεί να αυξήσει τον όγκο της εξαχθείσας από τη συλλογή κειμένων πληροφορίας. Πρώτα κατηγοριοποιείται ο όρος-στόχος σε μια από 20 προκαθορισμένες κατηγορίες, όπως «Αθλητής», «Συγγραφέας», «Θρησκεία» κ.ά. Οι Chu-Carroll κ.ά. αναφέρουν ότι η κατηγορία του όρου-στόχου προκύπτει μετά από συντακτική ανάλυση της ερώτησης και αναγνώριση των ονομάτων οντοτήτων που περιέχονται σε αυτήν, χωρίς όμως να εξηγούν περισσότερο αυτό το στάδιο. Σε κάθε κατηγορία όρων-στόχων αντιστοιχεί ένα σύνολο από προκαθορισμένες χειρωνακτικά κατασκευασμένες ερωτήσεις (π.χ. «Πότε γεννήθηκε ο/η όρος-στόχος;», αν ο όρος-στόχος έχει καταταγεί ως πρόσωπο), οι οποίες δίνονται στο σύστημα αναδρομικά. Οι κατηγορίες είναι οργανωμένες ιεραρχικά, ώστε για την κατηγορία «Αθλητής» να γίνονται και οι ερωτήσεις της ιεραρχικά ανώτερης κατηγορίας «Πρόσωπο». Υποθέτουμε ότι τα αποτελέσματα των αναδρομικών ερωτήσεων προστίθενται στην απάντηση, αλλά οι Chu-Carroll κ.ά. δεν παρέχουν περισσότερες λεπτομέρειες επ' αυτού.

2.4.5 Han κ.ά. – Πανεπιστήμιο της Κορέας

Το πρώτο βήμα στην προσέγγιση των Han κ.ά. [HaSo06] [HaCh04] είναι η ανάλυση της ερώτησης, κατά την οποία εντοπίζεται η λέξη-κεφαλή (head word) του όρου-στόχου και ο τύπος του (πρόσωπο, οργανισμός ή άλλο). Οι Han κ.ά. δεν διευκρινίζουν τι ακριβώς θεωρούν ως λέξη-κεφαλή του όρου-στόχου, αλλά αναφέρουν πως εντοπίζουν τη λέξη-κεφαλή μέσω συντακτικής ανάλυσης. Αναφέρουν, επίσης, πως εντοπίζουν τον τύπο του όρου-στόχου χρησιμοποιώντας ένα σύστημα αναγνώρισης ονομάτων οντοτήτων (named entity recognizer). Ο τύπος του όρου-στόχου χρησιμεύει σε επόμενα στάδια.

Για κάθε όρο-στόχο συλλέγουν κατόπιν κείμενα σε δύο στάδια. Στο πρώτο αναζητούν έγγραφα που περιέχουν τον όρο-στόχο. Αναφέρουν πως συλλέγουν έγγραφα χρησιμοποιώντας ως ερώτηση προς κάποια μηχανή αναζήτησης (δεν την προσδιορίζουν) τον όρο-στόχο (έχοντας αφαιρέσει τα stop-words), αλλά και ερωτήσεις-φράσεις (phrasal queries) αποτελούμενες από γειτονικές λέξεις του όρου-στόχου που ξεκινούν με κεφαλαία γράμματα, χωρίς να παρέχουν περισσότερες πληροφορίες επ' αυτού. Από τα ανακτηθέντα έγγραφα, εξάγουν ως αποσπάσματα τις προτάσεις που περιέχουν τη λέξη-κεφαλή του όρου-στόχου. Στη συνέχεια, σε ένα δεύτερο στάδιο, προσπαθούν να επεκτείνουν κάθε απόσπασμα, ώστε να περιλαμβάνει και τις γειτονικές του προτάσεις που περιέχουν κάποια αναφορά στον όρο-στόχο. Προκειμένου να επιλύσουν τις αναφορικές εκφράσεις (anaphora resolution), χρησιμοποιούν μια απλοϊκή τεχνική: θεωρούν ότι μια αναφορική έκφραση εντός μιας γειτονικής πρότασης αναφέρεται στον όρο-στόχο, αν η αναφορική έκφραση είναι το αντικείμενο της γειτονικής πρότασης και ο όρος-στόχος είναι το αντικείμενο της πρότασης αμέσως πριν τη γειτονική.

Κατόπιν το σύστημα προσπαθεί να ταιριάξει τα ανακτηθέντα αποσπάσματα με μια σειρά από χειρωνακτικά κατασκευασμένα συντακτικά πρότυπα (syntactic patterns), προκειμένου να εντοπιστούν ονοματικές και ρηματικές φράσεις που θα αποτελέσουν υποψήφιες απαντήσεις. Τα χρησιμοποιούμενα συντακτικά πρότυπα χωρίζονται γενικά σε 5 είδη, ανάλογα με τις ονοματικές ή ρηματικές φράσεις που εντοπίζουν:

- Ονοματικές φράσεις που έχουν άμεση συντακτική σχέση με τον όρο-στόχο (π.χ. «Former world and Olympic champion Alberto Tomba…», όπου ο όρος-στόχος είναι «Alberto Tomba»).
- Ονοματικές φράσεις που ακολουθούν το ρήμα «be» (π.χ. «TB is a bacterial disease»). Υποθέτουμε ότι ο όρος-στόχος πρέπει να είναι το υποκείμενο της πρότασης, χωρίς οι Han κ.ά. να το αναφέρουν αυτό.
- Ρηματικές φράσεις αναφορικών προτάσεων που εισάγονται με αναφορικές αντωνυμίες οι οποίες τροποποιούν (με τη συντακτική έννοια) απευθείας τον όρο-στόχο (π.χ. «Copland, who was born in Brooklyn, …», όπου ο όρος-στόχος είναι «Copland»).
- Φράσεις που εισάγονται με μετοχές ενεστώτα ή αόριστου (present or past participles) και τροποποιούν άμεσα τον όρο-στόχο ή το κύριο ρήμα που συνδέεται άμεσα με τον όρο-στόχο

(π.χ. «Tomba, known as La Bomba (the Bomb) for his explosive skiing style, had…», όπου ο όρος-στόχος είναι «Tomba»).

- Ρηματικές φράσεις των οποίων το υποκείμενο είναι ο όρος-στόχος (π.χ. «Iqra will initially broadcast eight hours a day», όπου ο όρος-στόχος είναι «Iqra»), εξαιρώντας ρηματικές φράσεις των οποίων το ρήμα περιλαμβάνεται σε μια λίστα ρημάτων που οι Han κ.ά. θεωρούν ότι δεν παρέχουν χρήσιμες πληροφορίες (π.χ. «be», «say», «talk», «tell»).

Οι Han κ.ά. χρησιμοποιούν το συντακτικό αναλυτή Conexor FDG parser [TaJa97]. Επειδή αυτός κάνει λάθη, επιστρατεύουν και συμπληρωματικές τεχνικές εντοπισμού λαθών, μερικές από τις οποίες βασίζονται σε ένα σύστημα εντοπισμού μερών του λόγου (POS tagger). Συγκεκριμένα, αν μια λέξη ανάμεσα στη πρώτη και την τελευταία κάποιας φράσης δεν συμπεριληφθεί από το συντακτικό αναλυτή στη φράση, την προσθέτουν. Αν η τελευταία λέξη μιας φράσης που επέστρεψε ο συντακτικός αναλυτής είναι επίθετο, προσδιορισμός ή πρόθεση, προστίθεται και η επόμενη ονοματική φράση. Τέλος, αν η τελευταία λέξη είναι σύνδεσμος ή αναφορική αντωνυμία, αφαιρείται από τη φράση.

Πριν αξιολογηθούν, οι φράσεις (υποψήφιες απαντήσεις) που εντοπίστηκαν συγκρίνονται μεταξύ τους, προς αποφυγή πλεονασμού. Αν δύο φράσεις έχουν περισσότερες από 70% κοινές λέξεις, τότε διαγράφεται η μία. (Οι Han κ.ά. δεν διευκρινίζουν τον τρόπο με τον οποίο επιλέγεται αυτή που θα διαγραφεί.) Αν οι κοινές τους λέξεις είναι λιγότερες από 30%, τότε δεν διαγράφεται καμία. Διαφορετικά ελέγχονται οι σημασιολογικές τάξεις των ουσιαστικών (αν είναι ονοματικές φράσεις) ή των ρημάτων (αν είναι ρηματικές). Τις σημασιολογικές τάξεις τις λαμβάνουν από το WordNet[6]. Αν οι τάξεις είναι ίδιες στις δύο φράσεις, πάλι διαγράφεται η μια φράση.

Οι φράσεις που απέμειναν, αξιολογούνται βάσει των παρακάτω μέτρων. Σε κάθε φράση αντιστοιχεί ένας «βαθμός πλεονασμού», που ισούται με το πλήθος των φράσεων που διαγράφηκαν λόγω σύγκρισης με αυτήν.

Πλεονασμός Κεφαλής (Head Redundancy)

$$Rdd(C) = \exp\left(\frac{r}{n}\right) - 1$$

Όπου:

C: η φράση που ελέγχεται

r: ο βαθμός πλεονασμού της φράσης

n: ο συνολικός αριθμός των υποψήφιων απαντήσεων/φράσεων

Το παραπάνω μέτρο επιβραβεύει φράσεις που οδήγησαν στη διαγραφή πολλών παρόμοιων φράσεων.

[6] Βλ. http://wordnet.princeton.edu/

Τοπικά Στατιστικά Όρων (Local Term Statistics)

$$Loc(C) = \frac{\sum_{t_i \in C} \frac{sf_i}{\max sf}}{|C|}$$

Όπου:

C: η φράση που ελέγχεται

t_i: μια λέξη που εμφανίζεται στη φράση

sf_i: το πλήθος των προτάσεων των ανακτηθέντων αποσπασμάτων στις οποίες εμφανίζεται η λέξη t_i

maxsf: το μέγιστο sf_i μεταξύ όλων των λέξεων t_i

|C|: το πλήθος των λέξεων περιεχομένου (content word, δηλαδή ρήματα, ουσιαστικά και επίθετα) της C.

Το παραπάνω μέτρο επιβραβεύει φράσεις που περιέχουν πολλές λέξεις οι οποίες είναι συχνές στα ανακτηθέντα αποσπάσματα.

Για τα δύο παρακάτω μέτρα συλλέγονται ορισμοί από εγκυκλοπαίδειες και γλωσσάρια. Για το μέτρο εξωτερικών ορισμών, οι ορισμοί που συλλέγονται αφορούν τον όρο-στόχο, ενώ για το μέτρο της ορολογίας ορισμών συλλέγονται ορισμοί για διάφορους όρους, ανάλογα με τον τύπο του όρου-στόχου (πρόσωπο, οργανισμός ή άλλο).

Εξωτερικοί Ορισμοί (External Definitions)

$$Ext(C) = \log\left(\frac{P(C|E)}{P(C)} + 1\right)$$

$$P(C|E) = \prod_{t_i \in C}\left(\frac{freq_{E_i}}{|E|}\right)^{\frac{1}{|C|}} \qquad P(C) = \prod_{t_i \in C}\left(\frac{freq_{B_i}}{|B|}\right)^{\frac{1}{|C|}}$$

Όπου:

C: η φράση που ελέγχεται

P(C|E): η πιθανότητα το C να είναι ένας από τους εξωτερικούς ορισμούς E

$freq_{Ei}$: το πλήθος των εμφανίσεων του t_i στους εξωτερικούς ορισμούς E για τον όρο-στόχο

$freq_{Bi}$: το πλήθος των εμφανίσεων του t_i στο σύνολο των φράσεων μιας γενικής συλλογής κειμένων B

Το παραπάνω μέτρο επιβραβεύει φράσεις που περιέχουν λέξεις που είναι συχνές στους εξωτερικούς ορισμούς του όρου-στόχου και σπάνιες σε γενικά κείμενα.

Ορολογία Ορισμού (Definition Terminology)

$$Tmn(C) = \frac{\sum_{t_i \in C} \log\left(\frac{P_D(t_i)}{P(t_i)} + 1\right)}{|C|}$$

Όπου:

$P_D(t)$: είναι η πιθανότητα το t να εμφανιστεί στους γενικούς εξωτερικούς ορισμούς D του τύπου του όρου-στόχου

P(t): είναι η πιθανότητα το t να εμφανιστεί σε μια γενική συλλογή κειμένων

Το παραπάνω μέτρο επιβραβεύει φράσεις που περιέχουν λέξεις οι οποίες είναι συχνές σε ορισμούς του τύπου του όρου-στόχου.

Τέλος, όλες οι φράσεις ταξινομούνται με βάση ένα γραμμικό συνδυασμό των παραπάνω μέτρων. Η τελική απάντηση σχηματίζεται επιλέγοντας τις κορυφαίες L φράσεις αυτής της λίστας, όπου L το μέγιστο μέγεθος απάντησης.

2.4.6 Xu Jun κ.ά. – Πανεπιστήμιο της Nankai

Στόχος της μεθόδου των Xu Jun κ.ά. [XuCa05] είναι να δημιουργήσει μια βάση με ορισμούς για κάθε όρο που εμφανίζεται σε μια δεδομένη συλλογή κειμένων. Ο χρήστης μπορεί μετά να θέσει ερωτήματα στη βάση και να λάβει τους αντίστοιχους ορισμούς, αν αυτοί περιλαμβάνονταν στην συλλογή κειμένων.

Σε πρώτο στάδιο εξάγονται από την συλλογή κειμένων παράγραφοι που ενδέχεται να περιέχουν ορισμούς. Αυτό γίνεται ως εξής: με τη χρήση ενός συντακτικού αναλυτή [XuHu00] εντοπίζονται σε κάθε πρόταση των παραγράφων οι βασικές ονοματικές φράσεις, δηλαδή οι ονοματικές φράσεις που δεν περιέχουν άλλες ονοματικές φράσεις. Για παράδειγμα στην πρόταση «Measures of manufacturing activity fell more than the overall measures» οι βασικές ονοματικές φράσεις είναι οι [Measures], [manufacturing activity] και [the overall measures]. Υποψήφιοι ορισμοί θεωρούνται όλες οι παράγραφοι που κάποια πρότασή τους ταιριάζει με ένα από τα παρακάτω πρότυπα:

- *όρος-στόχος* is a|an|the *
- *όρος-στόχος*, *, a|an|the *
- *όρος-στόχος* is one of *

όπου όρος-στόχος θεωρείται κάθε βασική ονοματική φράση που εμφανίζεται πρώτη σε μια πρόταση. Στον όρο-στόχο μπορεί να συμπεριληφθεί και μια δεύτερη βασική ονοματική φράση, αν αυτή χωρίζεται από την πρώτη με τη λέξη «for» ή «of».

Σε δεύτερο στάδιο χρησιμοποιείται μια Μ.Δ.Υ. παλινδρόμησης (SVM Regression [HeGr99]) εκπαιδευμένη σε χειρωνακτικά επισημειωμένα παραδείγματα για να καταταγούν οι υποψήφιοι ορισμοί (παράγραφοι) σε τρεις κατηγορίες: καλοί, μέτριοι και κακοί ορισμοί. Οι Xu Jun κ.ά. θεωρούν ότι ένας ορισμός είναι καλός αν σε αυτόν εμφανίζεται η γενική ιδέα του όρου-στόχου και τα χαρακτηριστικά του. Αντιθέτως, κακοί ορισμοί θεωρούνται αυτοί που δεν περιέχουν ούτε τη γενική ιδέα ούτε χαρακτηριστικά και μέτριοι θεωρούνται οι υπόλοιποι. Εναλλακτικά, χρησιμοποιείται μια Μ.Δ.Υ. κατάταξης (classification SVM) προκειμένου να καταταγούν οι υποψήφιοι ορισμοί ως καλοί ή κακοί. Τα πειράματα των Xu Jun κ.ά. έδειξαν ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των

δύο Μ.Δ.Υ.

Και στις δύο περιπτώσεις (παλινδρόμηση ή κατάταξη), οι ιδιότητες που χρησιμοποιούνται στη Μ.Δ.Υ. είναι οι παρακάτω:

- Ο όρος-στόχος εμφανίζεται στην αρχή της παραγράφου ή όχι.
- Ο όρος-στόχος αρχίζει με τις λέξεις «the», «a» ή «an» ή όχι.
- Όλες οι λέξεις στον όρο-στόχο αρχίζουν με κεφαλαία γράμματα ή όχι.
- Η παράγραφος περιέχει κάποια από τις λέξεις που οι Xu Jun κ.ά. έχουν ορίσει χειρωνακτικά ως αρνητικές (π.χ. «he», «she», «said») ή όχι.
- Ο όρος-στόχος περιέχει αντωνυμίες ή όχι.
- Ο όρος-στόχος περιέχει τη λέξη «of», «for» , «and», «or» ή «,» ή όχι.
- Ο όρος-στόχος εμφανίζεται περισσότερες από μία φορές στη παράγραφο ή όχι.
- Ο όρος-στόχος ακολουθείται από τις φράσεις «is a», «is an» ή «is the» ή όχι.
- Το πλήθος των προτάσεων στη παράγραφο.
- Το πλήθος των λέξεων στη παράγραφο.
- Το πλήθος των επιθέτων στη παράγραφο.
- Μια λέξη του κεντροειδούς ακολουθεί τον όρο-στόχο μέσα σε απόσταση Ν λέξεων ή όχι.

Η τελευταία ιδιότητα είναι αποτέλεσμα ενός απλού κεντροειδούς. Από όλες τις παραγράφους εκπαίδευσης συλλέγονται οι Ν λέξεις (κάθε μίας παραγράφου) που ακολουθούν τον όρο-στόχο και από αυτές εντοπίζονται οι πιο συχνές, οι οποίες και αποτελούν το κεντροειδές. Αν στη παράγραφο που αξιολογεί η Μ.Δ.Υ. μια από τις λέξεις του κεντροειδούς ακολουθεί τον όρο-στόχο μέσα σε απόσταση Ν λέξεων, η τελευταία ιδιότητα παίρνει την τιμή 1, διαφορετικά 0.

Τέλος, ελέγχονται οι ορισμοί και αφαιρούνται αυτοί που είναι πολύ όμοιοι μεταξύ τους, όπως στην ενότητα 2.3.2. Η ομοιότητα δύο ορισμών υπολογίζεται με βάση το πλήθος των τροποποιήσεων (εισαγωγές, διαγραφές κλπ.) που χρειάζονται για να μετασχηματιστεί ο ένας στον άλλον (Edit Distance).

Οι Xu Jun κ.ά. εκτέλεσαν, επίσης, πειράματα στα οποία οι υποψήφιοι ορισμοί ήταν προτάσεις, αντί για παράγραφοι. Στην περίπτωση αυτή αναφέρουν ότι χρησιμοποίησαν πρόσθετες ιδιότητες (π.χ. τη θέση της προτάσεως που αξιολογείται ως υποψήφιος ορισμός στην παράγραφό της). Και στις δύο περιπτώσεις (προτάσεις ή παράγραφοι ως υποψήφιοι ορισμοί), τα πειραματικά αποτελέσματα (για ερωτήσεις ορισμού) ήταν σημαντικά καλύτερα από τα αποτελέσματα του συστήματος ανάκτησης πληροφοριών Okapi [RoWa95].

3. ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΩΝ

3.1 Δεδομένα εκπαίδευσης και αξιολόγησης

Προκειμένου να αξιολογήσουμε τις μεθόδους αυτόματης επισημείωσης παραδειγμάτων εκπαίδευσης του προηγούμενου κεφαλαίου, επιλέξαμε 100 όρους-στόχους από το ευρετήριο μιας ηλεκτρονικής εγκυκλοπαίδειας, φροντίζοντας οι όροι-στόχοι να καλύπτουν μεγάλο εύρος θεμάτων, από ιατρικούς όρους και ιστορικά πρόσωπα μέχρι όρους φυσικής και όρους της καθημερινής ομιλίας.[7] Εξαγάγαμε, επίσης, από ιστοσελίδες που επέστρεψε η μηχανή αναζήτησης Altavista τα 5000 παράθυρα που αντιστοιχούσαν στους 100 όρους-στόχους (10 έγγραφα ανά όρο-στόχο και 5 παράθυρα ανά έγγραφο ή λιγότερα, αν δεν υπήρχαν τόσα), εξαιρώντας ιστοσελίδες που προέρχονταν από ηλεκτρονικές εγκυκλοπαίδειες. Από αυτά τα παράθυρα επιλέξαμε τυχαία 400, τα οποία επισημειώσαμε χειρωνακτικά ως ορισμούς ή μη-ορισμούς. Σημειώνουμε ότι προκαταρκτικά πειράματα που πραγματοποιήσαμε χρησιμοποιώντας 160 παράθυρα, που επιλέξαμε επίσης τυχαία ακολουθώντας την ίδια διαδικασία, έδειξαν ότι δύο άνθρωποι-κριτές (ο γράφων και ένας συνεργάτης του που είχε λάβει οδηγίες από τον πρώτο και μερικά παραδείγματα επισημείωσης) συμφωνούσαν σε μεγάλο βαθμό (K = 0,876, βλ. [EuGl04] για το μέτρο K) μεταξύ τους ως προς τις κατηγορίες των παραθύρων (ορισμοί ή μη-ορισμοί). Οπότε στη συνέχεια η χειρωνακτική επισημείωση παραθύρων έγινε από μόνο έναν άνθρωπο-κριτή (το γράφοντα). Συλλέξαμε, ακόμη, τους ορισμούς των 100 όρων-στόχων αξιολόγησης από εγκυκλοπαίδειες, όπως τους επέστρεψε η λειτουργία «define» της μηχανής αναζήτησης Google. Οι συνολικοί ορισμοί εγκυκλοπαιδειών που συλλέξαμε ήταν συνολικά 952.

Προκειμένου να εκπαιδεύσουμε τα συστήματα εντοπισμού ορισμών του προηγούμενου κεφαλαίου με τα οποία πειραματιστήκαμε, συλλέξαμε με τον ίδιο τρόπο, από το ευρετήριο της ηλεκτρονικής εγκυκλοπαίδειας, ως και 3.000 όρους-στόχους εξαιρώντας τους προηγούμενους 100. Συλλέξαμε, επίσης, τα αντίστοιχα παράθυρα ιστοσελίδων, τα οποία επισημειώσαμε αυτόματα· προέκυψαν έτσι ως 150.000 παράθυρα εκπαίδευσης (για 3.000 όρους-στόχους). Προκειμένου να αξιολογήσουμε κατόπιν τα συστήματα, συλλέξαμε με τον ίδιο τρόπο 200 πρόσθετους, νέους όρους-στόχους, καθώς και τα αντίστοιχα 10.000 παράθυρα ιστοσελίδων. Οι απαντήσεις των συστημάτων για τους 200 αυτούς όρους-στόχους αξιολογήθηκαν πάλι από έναν άνθρωπο-κριτή, το γράφοντα.

3.2 Πειράματα αυτόματης επισημείωσης παραδειγμάτων εκπαίδευσης

Στην ενότητα 2.2.2 παρουσιάσαμε τέσσερις διαφορετικές μεθόδους αυτόματης επισημείωσης παραδειγμάτων (παραθύρων) εκπαίδευσης. Αυτές είναι: η μέθοδος του Γιακουμή (δηλαδή του

7 Βλ. http://www.encyclopedia.com/. Η συγκεκριμένη ηλεκτρονική εγκυκλοπαίδεια περιλαμβάνει συχνά πολλούς ορισμούς ανά όρο, οι οποίοι προέρχονται από διαφορετικές εγκυκλοπαίδειες και λεξικά. Οι ορισμοί αυτοί θα μπορούσαν ενδεχομένως να προστεθούν σε εκείνους που λάβαμε μέσω της λειτουργίας «define» του Google, αν και ήταν γενικά πολύ λιγότεροι.

Γαλάνη, αλλά με ν-γράμματα), η μέθοδος του κεντροειδούς, το μέτρο ROUGE-W και η μέθοδος της δεύτερης, βοηθητικής Μ.Δ.Υ. Όλες οι μέθοδοι χρησιμοποιούν ένα άνω κατώφλι t_+ και ένα κάτω κατώφλι t_-, για να κρίνουν αν ένα παράθυρο εκπαίδευσης είναι ορισμός ή όχι. Για τους σκοπούς της αξιολόγησης των μεθόδων, θέτουμε $t_+ = t_-$ και μεταβάλλουμε την τιμή του κοινού πλέον κατωφλιού στο εύρος τιμών [0,1]. Για κάθε τιμή, υπολογίζουμε την ανάκληση και την ακρίβεια κάθε μεθόδου σύμφωνα με τους παρακάτω τύπους.

$$\text{Ακρίβεια}_{\text{ορισμών}} = \frac{TP}{TP + FP}$$

$$\text{Ανάκληση}_{\text{ορισμών}} = \frac{TP}{TP + FN}$$

$$\text{Ακρίβεια}_{\text{μη-ορισμών}} = \frac{TN}{TN + FN}$$

$$\text{Ανάκληση}_{\text{μη-ορισμών}} = \frac{TN}{TN + FP}$$

Όπου:

TP: Παράθυρα που κατατάσσονται σωστά ως ορισμοί

TN: Παράθυρα που κατατάσσονται σωστά ως μη-ορισμοί

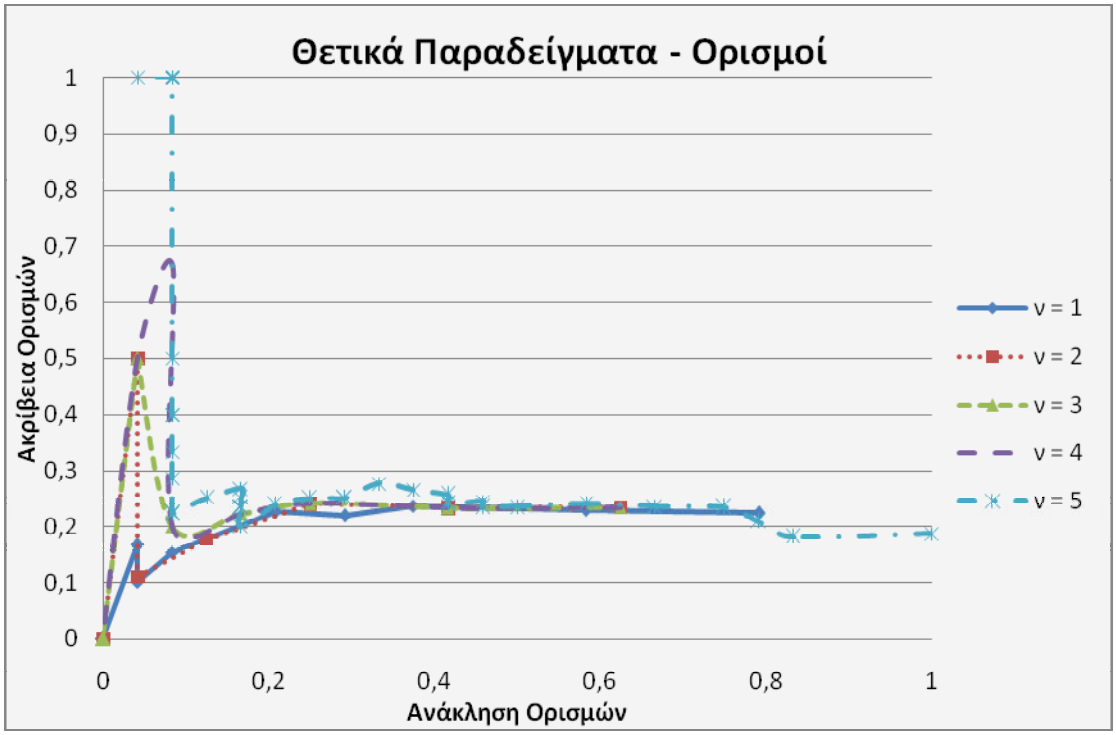
FP: Παράθυρα που κατατάσσονται λάθος ως ορισμοί

FN: Παράθυρα που κατατάσσονται λάθος ως μη-ορισμοί

Πριν συγκρίνουμε τις τέσσερις μεθόδους μεταξύ τους, πειραματιστήκαμε με κάθε μέθοδο ξεχωριστά, προκειμένου να επιλέξουμε τις καλύτερες τιμές των παραμέτρων της.

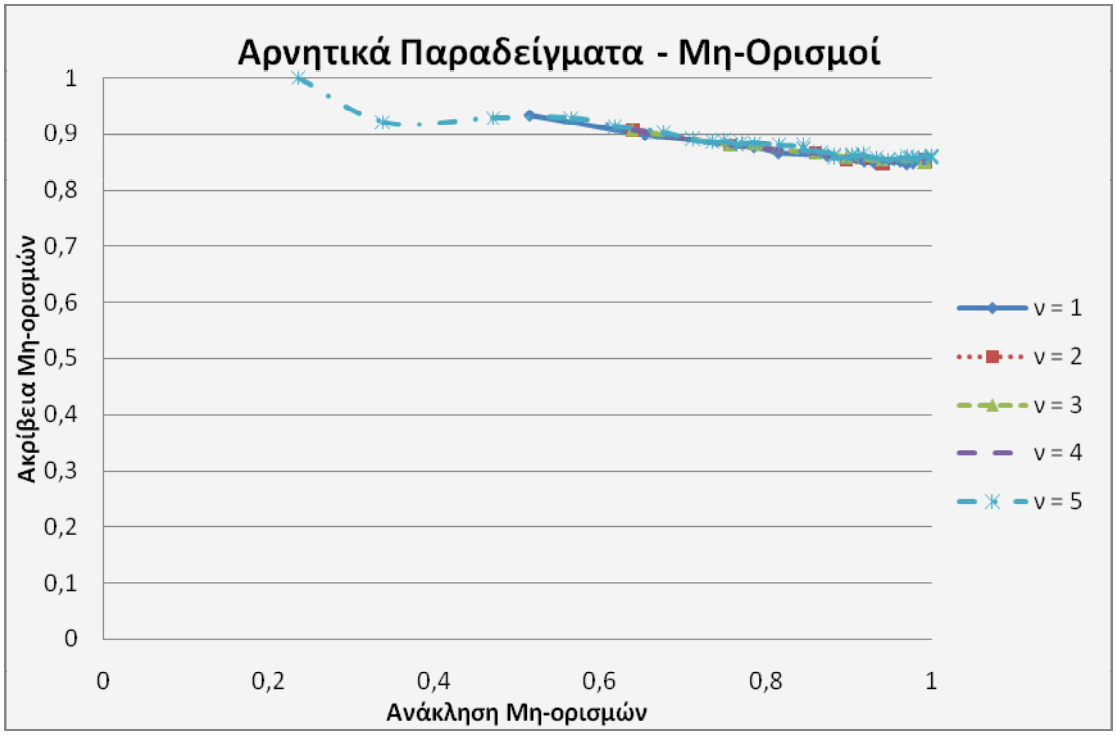
3.2.1 Επιλογή τιμών παραμέτρων της μεθόδου του Γιακουμή

Αρχικά κάναμε πειράματα με διάφορες τιμές της παραμέτρου ν (μήκος ν-γραμμάτων) της μεθόδου του Γιακουμή. Δοκιμάσαμε τις τιμές από 1 ως 5. Τα διαγράμματα που προέκυψαν είναι τα παρακάτω.



Εικόνα 1: Ακρίβεια και ανάκληση ορισμών της μεθόδου του Γιακουμή

Το πρώτο διάγραμμα μας δείχνει τη σχέση ακρίβειας και ανάκλησης της μεθόδου στην επισημείωση παραθύρων ως θετικά παραδείγματα, δηλαδή ως ορισμούς, για διαφορετικές τιμές του ν. Όπως βλέπουμε, τα καλύτερα αποτελέσματα εμφανίζονται για ν = 5, αν και η διαφορά στο διάγραμμα μεταξύ διαδοχικών τιμών του ν δεν είναι πολύ μεγάλη. Σημειώνουμε ότι για πολύ μικρή ανάκληση ορισμών, απομένουν ελάχιστα παράθυρα ορισμών, οπότε η ακρίβεια παίρνει ακραίες τιμές: για παράδειγμα, αν έχει απομείνει μόνο ένα παράθυρο ορισμού, παίρνει την τιμή 1 αν το παράθυρο είναι όντως ορισμός και τιμή 0 διαφορετικά.



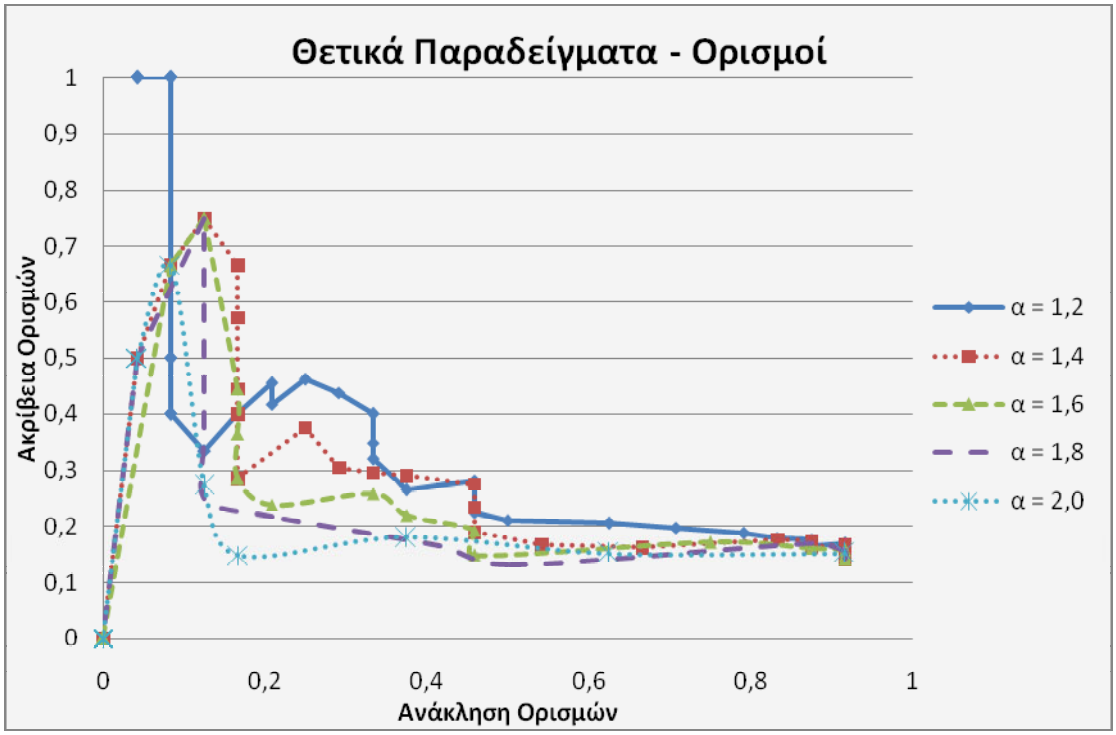
Εικόνα 2: Ακρίβεια και ανάκληση μη-ορισμών της μεθόδου του Γιακουμή

Στο αντίστοιχο διάγραμμα για τα αρνητικά παραδείγματα παρατηρούμε ότι δεν υπάρχει μεγάλη διαφορά μεταξύ των διαφορετικών τιμών του ν. Για όλες τις τιμές, η μέθοδος μπορεί με την ίδια ευκολία να αποφασίσει ότι ένα παράθυρο δεν είναι ορισμός. Επιλέξαμε την τιμή ν = 5, που οδηγεί σε καλύτερα αποτελέσματα στην κατηγορία των ορισμών.

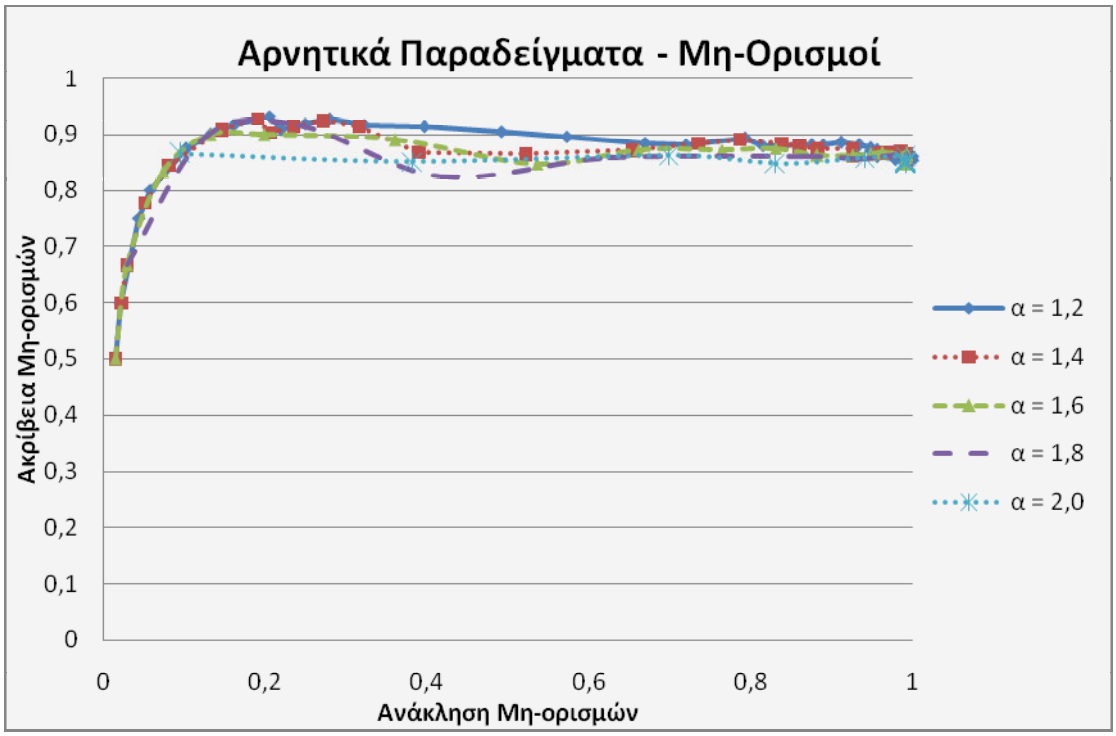
3.2.2 Επιλογή τιμών παραμέτρων της μεθόδου ROUGE-W

Τα αποτελέσματα της μεθόδου που χρησιμοποιεί το μέτρο ομοιότητας ROUGE-W εξαρτώνται από την τιμή της παραμέτρου α στη συνάρτηση βάρους $f(k) = k^{\alpha}$. Δοκιμάσαμε όλες τις τιμές του α μέσα στο εύρος (1,2] με βήμα 0,2.[8]

[8] Πειραματιστήκαμε, επίσης, με άλλες παραλλαγές του μέτρου ROUGE, αλλά τα αντίστοιχα αποτελέσματα παραλείπονται χάριν συντομίας. Το ROUGE-W ήταν γενικά το καλύτερο.



Εικόνα 3: Ακρίβεια και ανάκληση ορισμών της μεθόδου ROUGE-W



Εικόνα 4: Ακρίβεια και ανάκληση μη-ορισμών της μεθόδου ROUGE-W

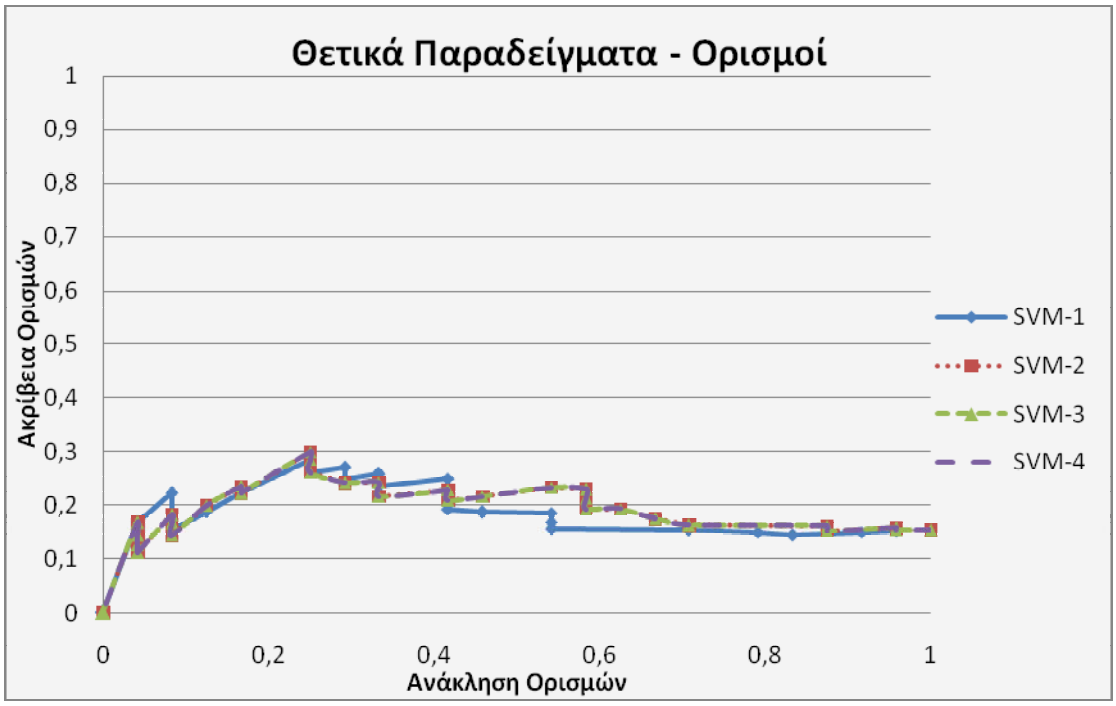
Παρατηρούμε πως τα καλύτερα αποτελέσματα εμφανίζονται εν γένει για τις τιμές α = 1,4 και α = 1,2. Επιλέξαμε την τιμή α = 1,4, αντί της α = 1,2, γιατί πετυχαίνει πολύ μεγαλύτερη ακρίβεια ορισμών για τιμές ανάκλησης ορισμών 0,1 ως και 0,2. Για μικρότερες τιμές ανάκλησης ορισμών υπερτερεί ως προς την ακρίβεια ορισμών η α = 1,4, αλλά παραμένουν ελάχιστα παράθυρα ορισμών, όπως προαναφέραμε. Παρατηρούμε, επίσης, πως όσο η τιμή του α γίνεται μεγαλύτερη του 1,4, τόσο

χειροτερεύουν τα αποτελέσματα. Φαίνεται πως από ένα σημείο και έπειτα, η αύξηση του βάρους που δίνεται στις διαδοχικές λέξεις της μέγιστης κοινής υποακολουθίας δύο παραθύρων επιδρά αρνητικά.

3.2.3 Επιλογή τιμών παραμέτρων της δεύτερης, βοηθητικής Μ.Δ.Υ.

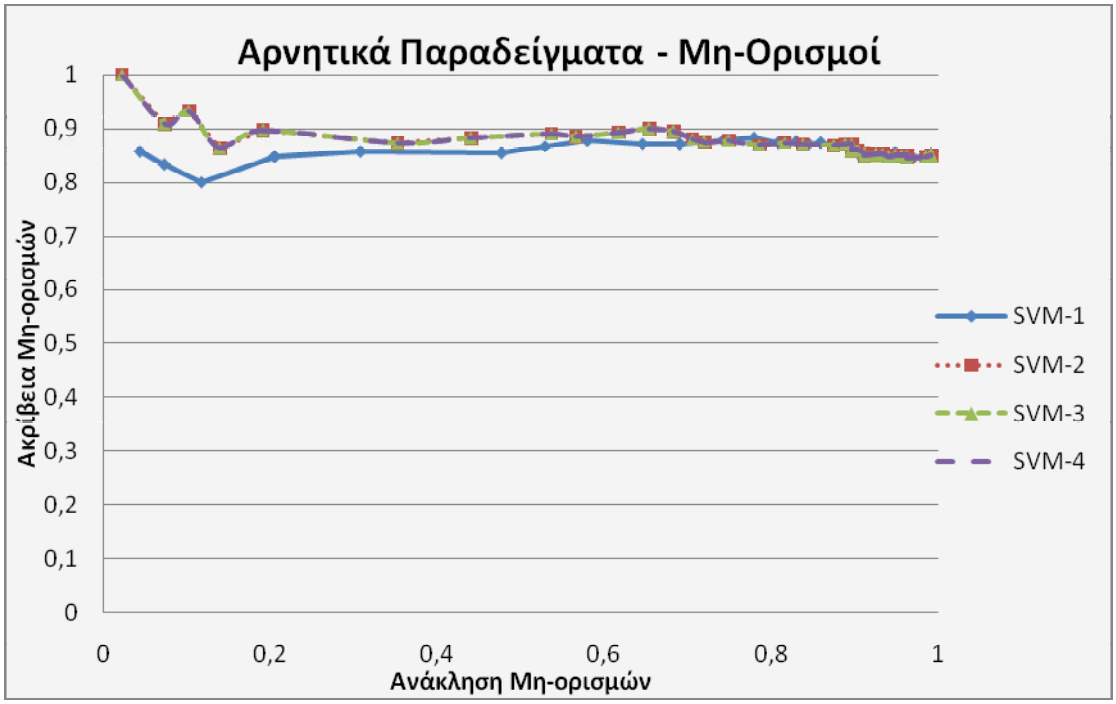
Υπενθυμίζουμε στον αναγνώστη ότι σε αυτή τη μέθοδο κάθε παράθυρο εκπαίδευσης μετατρέπεται σε ένα διάνυσμα που περιλαμβάνει τιμές μέτρων ομοιότητας, οι οποίες δείχνουν πόσο μοιάζει το παράθυρο με τους αντίστοιχους ορισμούς εγκυκλοπαιδειών. Το διάνυσμα δίνεται σε μια δεύτερη, βοηθητική Μ.Δ.Υ., που το επισημειώνει ως ορισμό ή μη-ορισμό, παράγοντας έτσι παραδείγματα εκπαίδευσης της κύριας Μ.Δ.Υ.

Πειραματιστήκαμε με τέσσερις συνδυασμούς ιδιοτήτων (μέτρα ομοιότητας) της βοηθητικής Μ.Δ.Υ. Ο πρώτος (SVM-1) περιέχει και τις 168 διαθέσιμες ιδιότητες (βλ. ενότητα 2.2.2.5). Ο δεύτερος (SVM-2) περιέχει μόνο τις ιδιότητες που προκύπτουν από τα απλά μέτρα (και όχι τις παραλλαγές τους για σύγκριση ν-γραμμάτων). Ο τρίτος (SVM-3) περιέχει μόνο τις ιδιότητες που προκύπτουν από τις παραλλαγές των μέτρων για σύγκριση ν-γραμμάτων (ν = 5). Ο τέταρτος (SVM-4) περιέχει μόνο τις ιδιότητες που προκύπτουν από τα απλά μέτρα για τα ζευγάρια 3, 4, 5 και 6 όπως αυτά περιγράφονται στην ενότητα 2.2.2.5.[9] Σε κάθε περίπτωση, χρησιμοποιήσαμε τη μέθοδο grid search που παρέχουν οι κατασκευαστές της υλοποίησης LibSVM (βλ. ενότητα 2.2.4), προκειμένου να επιλέξουμε τις υπόλοιπες τιμές των παραμέτρων της Μ.Δ.Υ. Η grid search επιλέγει παραμέτρους εκτελώντας διασταυρωμένη επικύρωση (cross-validation) στα δεδομένα εκπαίδευσης.



Εικόνα 5: Ακρίβεια και ανάκληση ορισμών της δεύτερης, βοηθητικής Μ.Δ.Υ.

[9] Δοκιμάσαμε και άλλους συνδυασμούς ιδιοτήτων. Αναφέρουμε εδώ μόνο τους πιο χαρακτηριστικούς. Ο SVM-4 ήταν εν γένει ο καλύτερος, με την έννοια που προαναφέρθηκε.



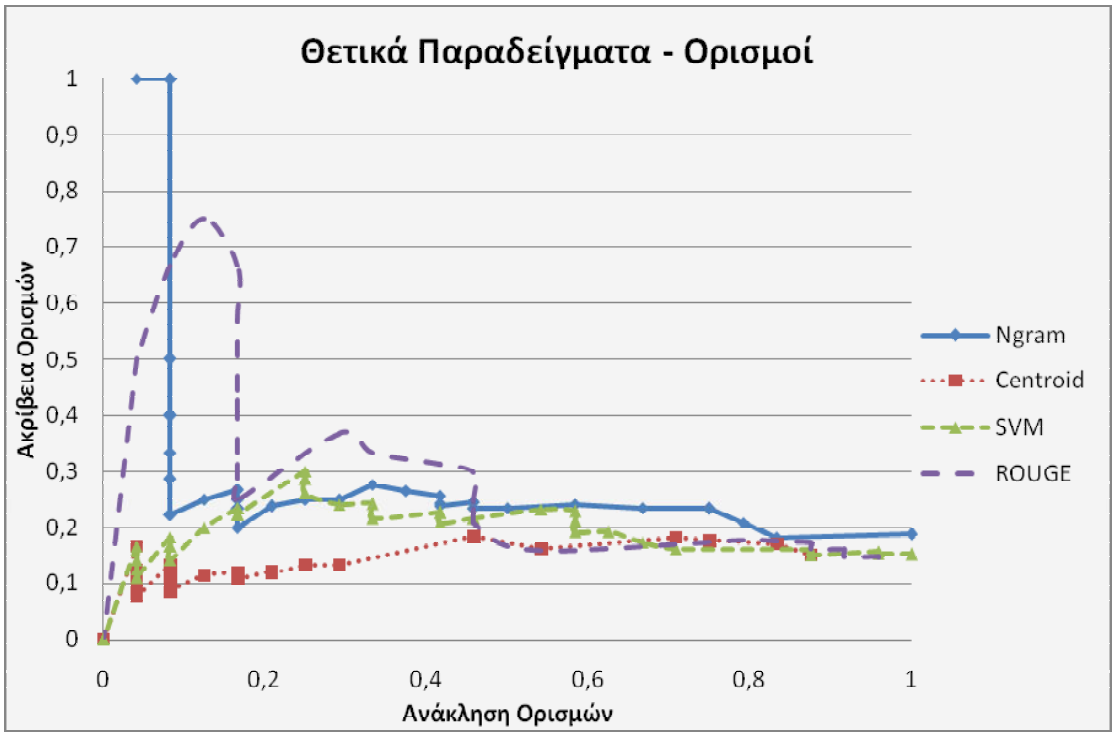
Εικόνα 6: Ακρίβεια και ανάκληση μη-ορισμών της δεύτερης, βοηθητικής Μ.Δ.Υ.

Οι συνδυασμοί SVM-2, SVM-3 και SVM-4 έχουν τα ίδια αποτελέσματα, τόσο στα θετικά όσο και στα αρνητικά παραδείγματα. Μια πιθανή εξήγηση είναι πως τα μέτρα έχουν πολύ παρόμοια αποτελέσματα τόσο στις κανονικές τους μορφές όσο και στις παραλλαγές τους για σύγκριση ν-γραμμάτων, κάτι που εξηγεί την ομοιότητα των αποτελεσμάτων των SVM-2 και SVM-3. Επιπλέον, η ομοιότητα των SVM-2 και SVM-3 με το SVM-4 δείχνει πως οι ιδιότητες που προκύπτουν από τα ζευγάρια 1 και 2 (οι οποίες αγνοούνται στο SVM-4) δεν προσφέρουν καμία πρόσθετη πληροφορία. Η χρήση όλων των διαθέσιμων ιδιοτήτων μαζί (SVM-1) χειροτερεύει τα αποτελέσματα, ιδιαίτερα στην περίπτωση των μη-ορισμών, κάτι που επίσης δείχει πως υπάρχει μεγάλος πλεονασμός μεταξύ των διαθέσιμων ιδιοτήτων. Μπορούμε, λοιπόν, να θεωρήσουμε ως καλύτερο το συνδυασμό ιδιοτήτων SVM-4, αφού πετυχαίνει τα ίδια (ή καλύτερα) αποτελέσματα με τους άλλους συνδυασμούς χρησιμοποιώντας λιγότερες (57) ιδιότητες.

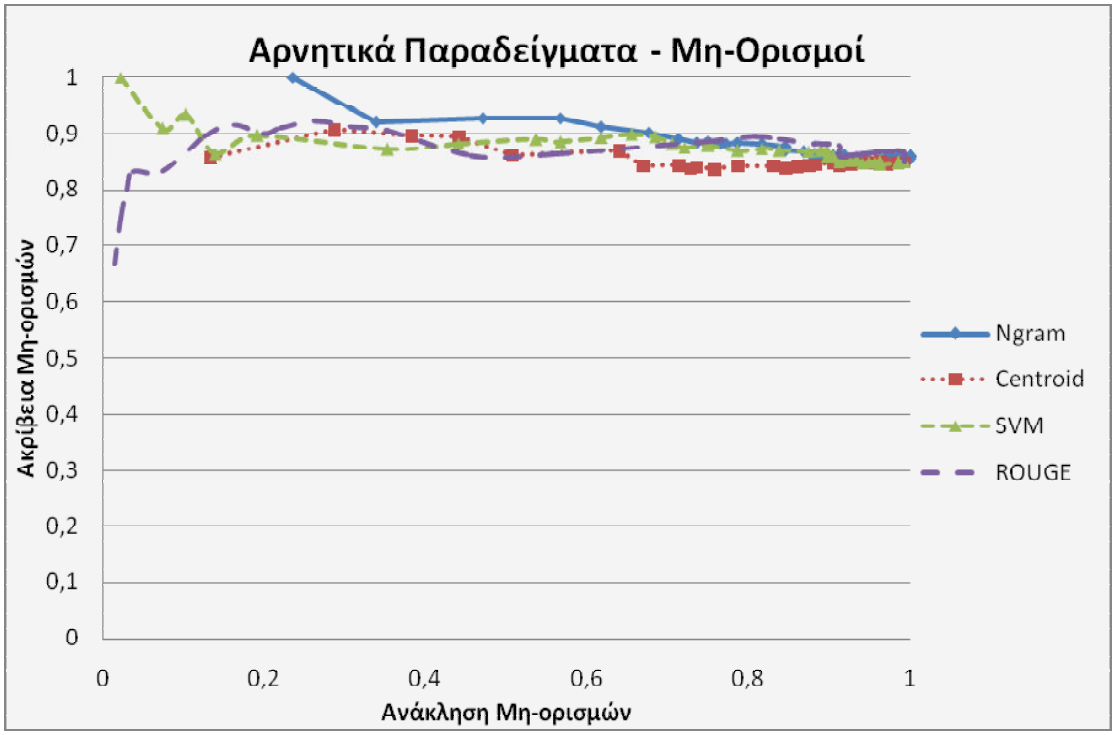
Ως μια προσπάθεια βελτίωσης των αποτελεσμάτων, δοκιμάσαμε να προσθέσουμε δύο ακόμα ιδιότητες στα διανύσματα της βοηθητικής Μ.Δ.Υ., οι οποίες έδειχναν τις «απόψεις» άλλων μεθόδων επισημείωσης παραδειγμάτων. Η πρώτη ιδιότητα ήταν το αποτέλεσμα της μεθόδου του Γιακουμή (με ν = 5) και η δεύτερη το αποτέλεσμα της μεθόδου του κεντροειδούς. Προέκυψαν έτσι τρεις νέοι συνδυασμοί ιδιοτήτων: ένας που περιείχε επιπλέον την ιδιότητα της μεθόδου του Γιακουμή, ένας που περιείχε επιπλέον την ιδιότητα του κεντροειδούς και ένας που περιείχε και τις δύο. Τα αποτελέσματα δεν επηρεάστηκαν πρακτικά καθόλου, όμως, από τις πρόσθετες ιδιότητες και τα παραλείπουμε χάριν συντομίας.

3.2.4 Σύγκριση μεθόδων αυτόματης επισημείωσης

Έχοντας επιλέξει τις τιμές των παραμέτρων των μεθόδων αυτόματης επισημείωσης παραδειγμάτων εκπαίδευσης, συγκρίναμε τις μεθόδους μεταξύ τους.



Εικόνα 7: Ακρίβεια και ανάκληση ορισμών των μεθόδων επισημείωσης παραδειγμάτων

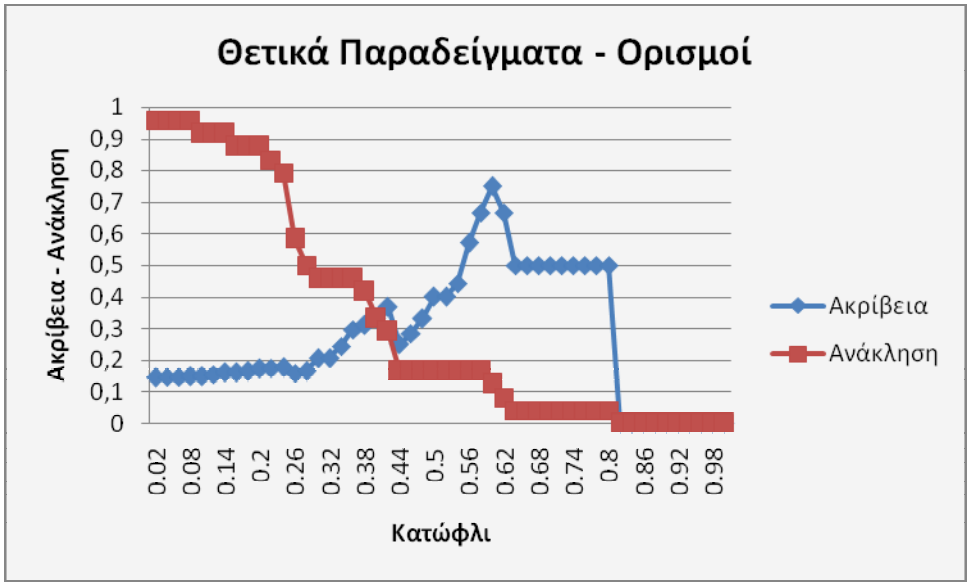


Εικόνα 8: Ακρίβεια και ανάκληση μη-ορισμών των μεθόδων επισημείωσης παραδειγμάτων

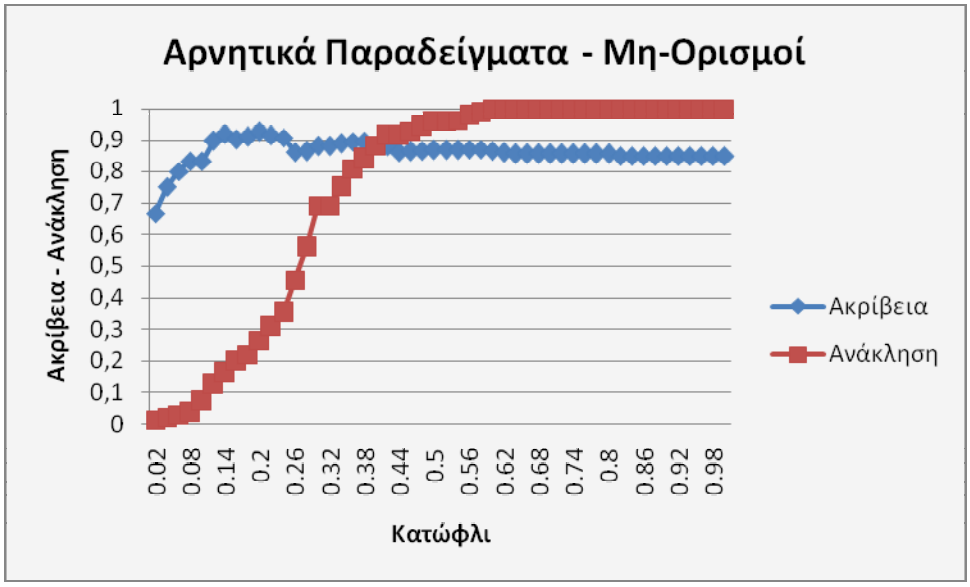
Από τα παραπάνω διαγράμματα φαίνεται πως σε γενικές γραμμές η καλύτερη μέθοδος είναι η χρήση του μέτρου ROUGE-W, αλλά η διαφορά από τη μέθοδο του Γιακουμή (ν-γράμματα) είναι μικρή. Χειρότερη είναι, όπως ήταν αναμενόμενο λόγω της απλότητάς της, η μέθοδος του κεντροειδούς, ενώ η βοηθητική Μ.Δ.Υ. δεν οδήγησε σε αποτελέσματα που να δικαιολογούν τη χρήση

της.

Έχοντας εντοπίσει την καλύτερη μέθοδο, μας μένει να προσδιορίσουμε τα κατώφλια t_+ και t_- που θα χρησιμοποιήσουμε στο σύστημα ερωταποκρίσεων για την εκπαίδευση της κύριας Μ.Δ.Υ. Στα παρακάτω διαγράμματα δείχνουμε την ανάκληση και ακρίβεια των δύο κατηγοριών (ορισμοί και μη-ορισμοί) συναρτήσει ενός κοινού κατωφλίου $t = t_+ = t_-$, για τη μέθοδο του ROUGE-W. Τελικά επιλέξαμε τις τιμές $t_- = 0,3$ και $t_+ = 0,58$. Με άλλα λόγια, θεωρούμε ένα παραδείγματα εκπαίδευσης ως θετικό, αν η ομοιότητά του (σύμφωνα με το ROUGE-W) με τους αντίστοιχους ορισμούς εγκυκλοπαιδειών είναι μεγαλύτερη από $t_+ = 0,58$. Αν η ομοιότητα είναι μικρότερη από $t_- = 0,3$, το παράδειγμα θεωρείται αρνητικό. Τα υπόλοιπα παραδείγματα εκπαίδευσης αγνορούνται. Για το t_+ επιλέξαμε την τιμή που οδηγεί στη μεγαλύτερη ακρίβεια ορισμών (0,66%), ώστε να έχουμε όσο το δυνατόν λιγότερα λανθασμένα θετικά παραδείγματα ορισμών, εις βάρος της ανάκλησης ορισμών (0,16%). Κατόπιν επελέγη η τιμή του t_- που διατηρεί την αρχική (πριν την αφαίρεση των παραδειγμάτων με ομοιότητα μεταξύ των t_- και t_+) αναλογία θετικών και αρνητικών παραδειγμάτων εκπαίδευσης· η τιμή αυτή οδήγησε σε ακρίβεια μη-ορισμών 0,7% και ανάκληση ορισμών 0,02%, τις οποίες θεωρήσαμε ικανοποιητικές. Γενικά, το μεγαλύτερο πρόβλημα είναι πως αναγκαζόμαστε να αγνοήσουμε ένα πολύ μεγάλο μέρος των παραδειγμάτων εκπαίδευσης που είναι στην πραγματικότητα ορισμοί (η ανάκληση ορισμών είναι μόλις 0,16%), προκειμένου να επιτύχουμε μια σχετικά (αλλά όχι απολύτως, μόνο 0,66%) ικανοποιητική ακρίβεια ορισμών. Αυτό ενδέχεται να έχει ως αποτέλεσμα η Μ.Δ.Υ. να συναντά κατά την εκπαίδευσή της μικρότερη ποικιλία ορισμών από ό,τι κατά τη χρήση της.



Εικόνα 9: Ακρίβεια και ανάκληση ορισμών της ROUGE-W για διαφορετικά κατώφλια



Εικόνα 10: Ακρίβεια και ανάκληση μη-ορισμών της ROUGE-W για διαφορετικά κατώφλια

3.3 Πειράματα συστημάτων ερωταποκρίσεων

3.3.1 Μέτρα αξιολόγησης

Ακολουθώντας τους κανονισμούς των διαγωνισμών TREC 2000 [Vo00] και TREC 2001 [Vo01] για τις ερωτήσεις ορισμού, το σύστημά μας επιστρέφει μια λίστα με τις πέντε απαντήσεις (παράθυρα του όρου-στόχου) που έκρινε ως πιθανότερο να περιέχουν αποδεκτό ορισμό του όρου-στόχου. Στην περίπτωση που τουλάχιστον ένα από αυτά τα παράθυρα περιέχει ορισμό, τότε θεωρούμε ότι το σύστημα κατάφερε και απάντησε σωστά.

Για να μπορούμε να κρίνουμε και το κατά πόσο το σύστημα επιστρέφει σωστές απαντήσεις στις υψηλές θέσεις της λίστας, χρησιμοποιούμε τη Μέση Αντίστροφη Κατάταξη (Mean Reciprocal Rank). Για να την υπολογίσουμε, αντιστοιχούμε σε κάθε ερώτηση ένα βαθμό. Αυτός ισούται με 1 δια τη θέση της πρώτης σωστής απάντησης στη λίστα (1 – 5). Αν δεν εμφανίζεται σωστή απάντηση, τότε ο βαθμός παίρνει την τιμή 0. Παίρνοντας τον μέσο όρο των βαθμών προκύπτει η αριθμητική τιμή της Μέσης Αντίστροφης Κατάταξης. Αυτή κυμαίνεται από 0 έως 1, και όσο υψηλότερη είναι τόσο λιγότερες απαντήσεις απαιτείται να επιστρέψει το σύστημα μέχρι να επιστραφεί η σωστή.

Τέλος, υπολογίζεται και το ποσοστό επιτυχίας του συστήματος αν του επιτραπεί να δώσει μόνο μία απάντηση ανά ερώτηση, αυτήν που θεωρεί ως πιο πιθανή. Ο σκοπός αυτής της μέτρησης είναι κυρίως η σύγκριση με μεθόδους προηγούμενων εργασιών.

3.2.2 Απλά συστήματα σύγκρισης

Για να δείξουμε ότι το σύστημα ερωταποκρίσεων που παρουσιάζουμε είναι καλύτερο από απλές μεθόδους, θα χρησιμοποιήσουμε τρία απλά συστήματα (baselines).

Τα δύο πρώτα, που θα ονομάσουμε Baseline-1 και Baseline-2, δεν χρησιμοποιούν μηχανική μάθηση. Το πρώτο απλά επιστρέφει ως απαντήσεις τα πρώτα παράθυρα του όρου-στόχου των πέντε κορυφαίων ιστοσελίδων (ένα παράθυρο από κάθε ιστοσελίδα) που επέστρεψε η μηχανή αναζήτησης. Το δεύτερο επιλέγει τυχαία 5 παράθυρα του όρου-στόχου από το σύνολο των 50 παραθύρων του όρου-στόχου που ανακτώνται από τις ιστοσελίδες που επέστρεψε η μηχανή αναζήτησης και τα επιστρέφει. Όταν επιτρέπεται μόνο μία απάντηση το Baseline-1 επιστρέφει το πρώτο παράθυρο της κορυφαίας ιστοσελίδας και το Baseline-2 ένα μόνο τυχαίο παράθυρο. Και οι δύο αυτές μέθοδοι δεν κάνουν καμία ανάλυση στον όρο-στόχο ή στα παράθυρα.

Το τρίτο απλό σύστημα, που ονομάζουμε Centroid, είναι ουσιαστικά η μέθοδος του κεντροειδούς, όπως αυτή παρουσιάζεται στην παράγραφο 2.2.2.3. Για κάθε όρο-στόχο, συλλέγονται όλες οι υποψήφιες απαντήσεις (παράθυρα) και κατασκευάζεται ένα κεντροειδές από αυτές. Κάθε υποψήφια απάντηση αξιολογείται βάσει της ομοιότητας συνημιτόνου της (cosine similarity) με το κεντροειδές και επιστρέφονται οι πέντε (ή μία) υποψήφιες απαντήσεις με τη μεγαλύτερη ομοιότητα.

Τα αποτελέσματα των τριών συστημάτων παρουσιάζονται στον παρακάτω πίνακα, από όπου προκύπτει ότι καλύτερη μεταξύ των τριών είναι η Centroid.

	Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)	**Ορθότητα όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
Baseline 1	9,548 %	31,156 %	0,164
Baseline 2	13,568 %	47,236 %	0,245
Centroid	14,573 %	51,256 %	0,263

3.3.3 Συστήματα τρίτων

Στην παράγραφο 2.4 παρουσιάσαμε έξι συστήματα άλλων ερευνητών. Δυστυχώς οι υλοποιήσεις των περισσοτέρων από αυτά τα συστήματα δεν είναι ελεύθερα διαθέσιμες. Επίσης, οι περιγραφές των μεθόδων τους δεν είναι αρκετά πλήρεις, ώστε να μπορέσουμε να τις υλοποιήσουμε. Ακόμη, πολλά από τα συστήματα χρησιμοποιούν ορισμούς ηλεκτρονικών εγκυκλοπαιδειών, ενώ ο σκοπός μας είναι να βρίσκουμε ορισμούς όρων που δεν περιλαμβάνονται σε εγκυκλοπαίδειες. Τελικά, από τα έξι συστήματα μπορέσαμε να πειραματιστούμε μόνο με αυτά των Cui κ.ά. και των Chu-Carroll κ.ά.

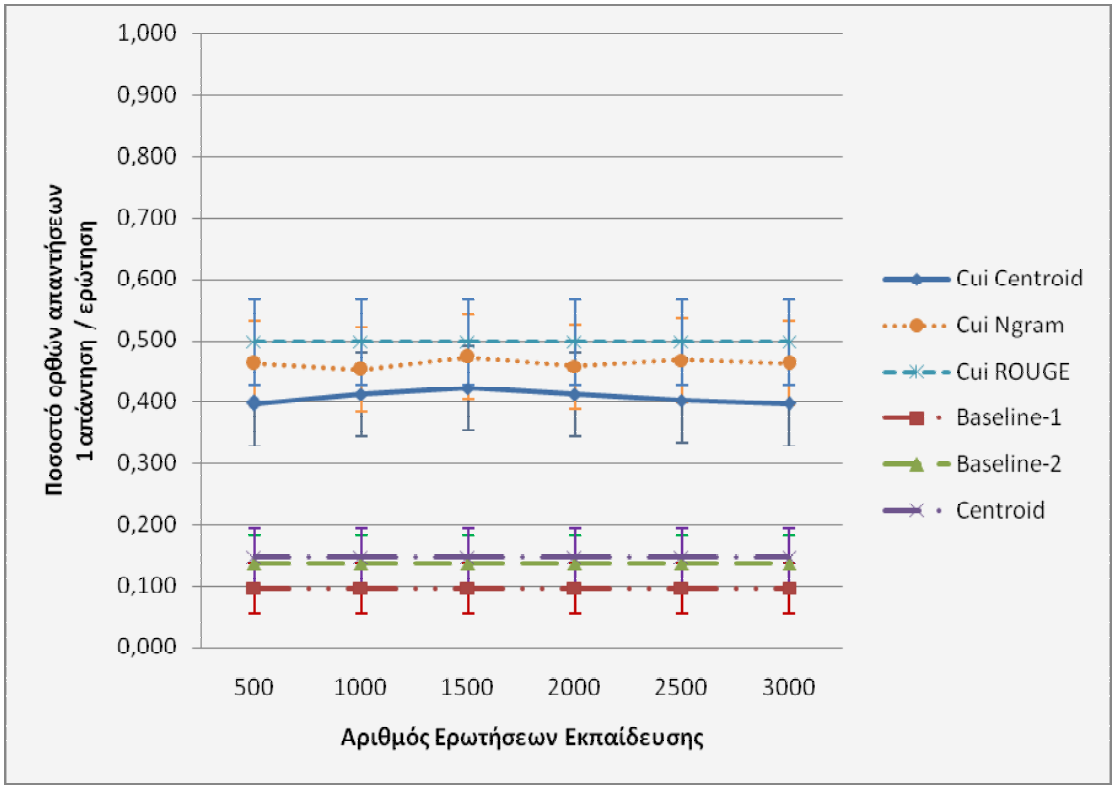
Για το σύστημα των Cui κ.ά. χρησιμοποιήσαμε την υλοποίηση που παρέχουν οι ίδιοι, αλλά δοκιμάσαμε επιπλέον να επισημειώνουμε τα παραδείγματα εκπαίδευσης χρησιμοποιώντας το μέτρο ROUGE-W (βλ. ενότητα 2.2.2.4) ή τη μέθοδο του Γιακουμή (βλ. ενότητα 2.2.2.2) και τους

αντίστοιχους ορισμούς εγκυκλοπαιδειών, αντί του κεντροειδούς που χρησιμοποιούν κανονικά οι Cui κ.ά. για αυτό το σκοπό (βλ. ενότητα 2.2.2.3).[10] Στο σύστημα των Cui κ.ά., όπως αναφέραμε στην παράγραφο 2.4.1, αφού εξαχθούν τα παράθυρα για κάποιον όρο-στόχο, κατασκευάζεται το κεντροειδές τους και συγκρίνονται όλα τα παράθυρα με το κεντροειδές μέσω του μέτρου απόστασης συνημιτόνου. (Εξακολουθούμε να χρησιμοποιούμε πάντα το κεντροειδές σε αυτό το στάδιο, ανεξαρτήτως του αν χρησιμοποιούμε το κεντροειδές, το μέτρο ROUGE-W ή τη μέθοδο του Γαλάνη κατά την επισημείωση των παραδειγμάτων εκπαίδευσης.) Τα 10 παράθυρα που είναι πιο κοντά στο κεντροειδές θεωρούνται πιθανότεροι ορισμοί και προωθούνται στα επόμενα στάδια του συστήματος.

Κάναμε πειράματα με το σύστημα των Cui κ.ά. για 500, 1000, 1500, 2000, 2500 και 3000 ερωτήσεις εκπαίδευσης, χρησιμοποιώντας για την αυτόματη επισημείωση των παραδειγμάτων εκπαίδευσης είτε τη μέθοδο του κεντροειδούς, που χρησιμοποιούν κανονικά οι Cui κ.ά., είτε τη μέθοδο του Γιακουμή, είτε το μέτρο ROUGE-W. Σε κάθε περίπτωση, επιτρέψαμε στο σύστημα να επιστρέφει μόνο μια απάντηση ανά ερώτηση. Στις παραμέτρους του συστήματος δώσαμε τις τιμές $\alpha = 0,6$ και $\lambda = 0,7$, στις οποίες καταλήξαμε κάνοντας δοκιμές με το αυθεντικό σύστημα των Cui κ.ά. στο εύρος [0,1] με βήμα 0,1 πάνω σε 160 παράθυρα που επιλέχθηκαν τυχαία με τη ίδια διαδικασία που περιγράψαμε στην ενότητα 3.1. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα και το διάγραμμα που ακολουθεί.

Ερωτήσεις Εκπαίδευσης	**Ορθότητα με κεντροειδές (Cui Centroid)**	**Ορθότητα με μέθοδο Γιακουμή (Cui NGram)**	**Ορθότητα με μέτρο ROUGE (Cui ROUGE)**
500	39,698 %	46,231 %	49,749 %
1000	41,206 %	45,226 %	49,749 %
1500	42,211 %	47,236 %	49,749 %
2000	41,206 %	45,729 %	49,749 %
2500	40,201 %	46,734 %	49,749 %
3000	39,698 %	46,231 %	49,749 %

[10] Το σύστημα των Cui κ.ά. διατίθεται από τη διεύθυνση http://www.cuihang.com/software.html.



Εικόνα 11: Ποσοστό ορθότητας του συστήματος των Cui κ.ά.

Τα αποτελέσματα φαίνεται να επιβεβαιώνουν τα συμπεράσματα της ενότητας 3.2.4, δείχνουν δηλαδή ότι η καλύτερη μέθοδος επισημείωσης παραδειγμάτων εκπαίδευσης είναι η χρήση του μέτρου ROUGE-W, με δεύτερη καλύτερη τη μέθοδο του Γιακουμή και χειρότερη τη χρήση του κεντροειδούς. Χρησιμοποιώντας, δηλαδή, τις μεθόδους ROUGE-W και Γιακουμή καταφέρνουμε να βελτιώσουμε τις επιδόσεις του αρχικού συστήματος των Cui κ.ά. Οι διαφορές, όμως, είναι μικρές· τα διαστήματα λάθους στο παραπάνω διάγραμμα είναι διαστήματα εμπιστοσύνης 95%.[11] Με οποιαδήποτε από τις τρεις μεθόδους επισημείωσης, πάντως, το σύστημα των Cui κ.ά. είναι αισθητά καλύτερο από τα απλά συστήματα και η διαφορά είναι στατιστικά σημαντική.

Θεωρούμε, επομένως, πως η καλύτερη επιλογή είναι η χρήση του μέτρου ROUGE-W στο σύστημα των Cui κ.ά. και αυτή θα χρησιμοποιήσουμε στις επόμενες συγκρίσεις· καλούμε Cui-ROUGE το σύστημα που προκύπτει. Τα αποτελέσματα για αυτή την περίπτωση και όταν το σύστημα επιτρέπεται να επιστρέφει πέντε παράθυρα ανά ερώτηση παρατίθενται στον παρακάτω πίνακα. Προσθέσαμε αποτελέσματα και για 50, 100 και 250 ερωτήσεις εκπαίδευσης.

Ερωτήσεις Εκπαίδευσης	**Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**	**Ορθότητα όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
50	46,231 %	82,412 %	0,604

[11] Λόγω χρονικών περιορισμών δεν εκτελέσαμε άλλους ελέγχους στατιστικής σημαντικότητας, αλλά ελπίζουμε να τους εκτελέσουμε στο άμεσο μέλλον.

100	50,754 %	82,915 %	0,630
250	49,246 %	84,925 %	0,633
500	49,749 %	87,940 %	0,654
1000	49,749 %	87,940 %	0,654
1500	49,749 %	87,940 %	0,654
2000	49,749 %	87,940 %	0,654
2500	49,749 %	87,940 %	0,654
3000	49,749 %	87,940 %	0,654

Παρατηρούμε ότι με 50 μόνο ερωτήσεις εκπαίδευσης το ποσοστό ορθότητας του συστήματος Cui-ROUGE είναι αρκετά υψηλό. Το σύστημα πετυχαίνει γρήγορα το μέγιστο ποσοστό ορθότητάς του με 500 ερωτήσεις εκπαίδευσης· το ποσοστό ορθότητας παραμένει κατόπιν σταθερό, ανεξαρτήτως του αριθμού ερωτήσεων εκπαίδευσης. Σημειώνουμε εδώ πως στα παραπάνω αποτελέσματα δεν έχει γίνει αφαίρεση πλεονασμού, αφού η υλοποίηση που έχουμε δεν τη περιλαμβάνει.

Στην περίπτωση του συστήματος PIQUANT των Chu-Caroll κ.ά., υλοποιήσαμε οι ίδιοι το συστατικό που σχετίζεται με ερωτήσεις ορισμού (βλ. ενότητα 2.4.4) βασιζόμενοι στα άρθρα τους. Η μόνη διαφορά από τη μέθοδο που προτείνουν είναι ότι δεν χρησιμοποιούμε τη μηχανή αναζήτησης JuruXML για τον εντοπισμό των κειμένων και την εξαγωγή των παραθύρων, αλλά τη μηχανή AltaVista όπως και στις άλλες μεθόδους μας. Τα αποτελέσματα του συστατικού του PIQUANT συνοψίζονται στον παρακάτω πίνακα. Παρατηρούμε ότι οι επιδόσεις του συστατικού είναι αισθητά χειρότερες από εκείνες του Cui-ROUGE.

Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)	**Ορθότητα όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
19,095 %	30,151%	0,226

3.3.4 Πειράματα με το σύστημα της εργασίας

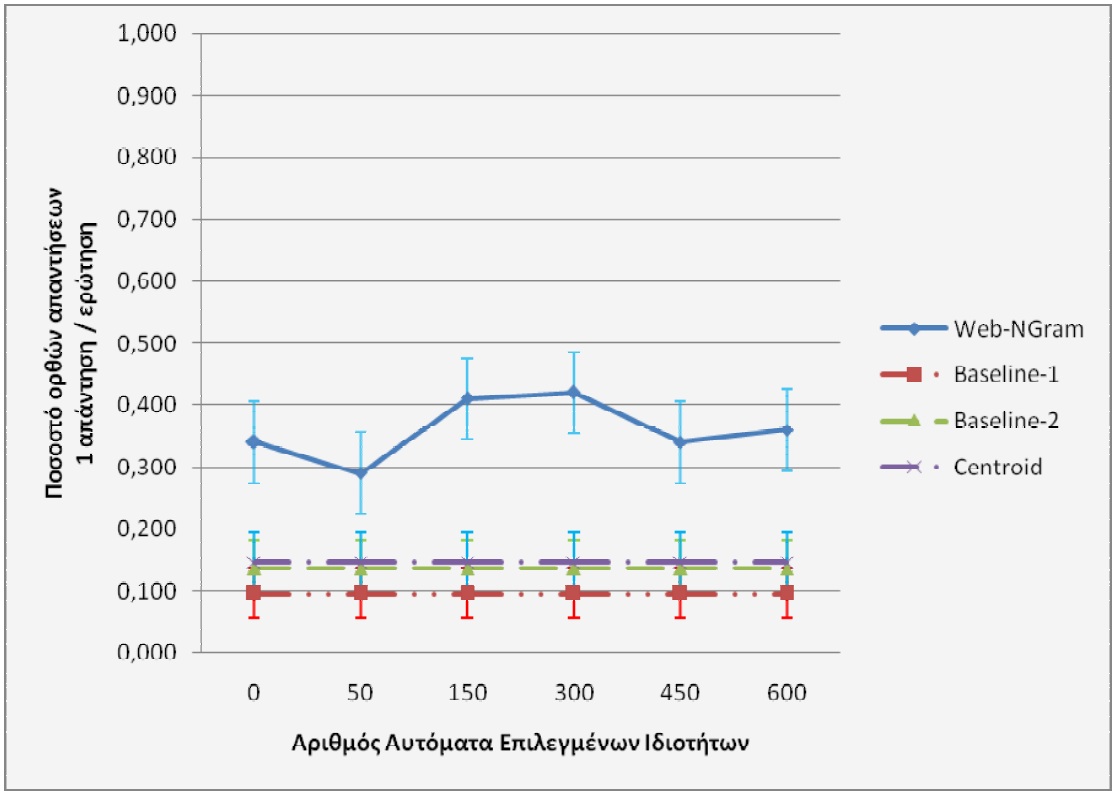
Όπως έχουμε ήδη αναφέρει, στην περίπτωση του δικού μας συστήματος κάθε παράθυρο μετατρέπεται σε ένα διάνυσμα που αποτελείται από 22 χειρωνακτικά επιλεγμένες ιδιότητες και έναν αριθμό αυτόματα επιλεγμένων ιδιοτήτων που αντιστοιχούν σε ν-γράμματα λέξεων πριν και μετά τον όρο-στόχο. Προκειμένου να επιλέξουμε το πλήθος των αυτόματα επιλεγμένων ιδιοτήτων, εκπαιδεύσαμε το σύστημα της εργασίας σε 2.000 ερωτήσεις εκπαίδευσης και το αξιολογήσαμε σε 100 ερωτήσεις διαφορετικές από όλες τις ερωτήσεις που χρησιμοποιήθηκαν σε άλλα πειράματα, χρησιμοποιώντας 0, 50, 150, 300, 450 και 600 αυτόματα επιλεγμένες ιδιότητες. Στη μελέτη αυτή, η επισημείωση των παραδειγμάτων εκπαίδευσης έγινε χρησιμοποιώντας τη μέθοδο του Γιακουμή (ν = 5, $t_- = 0{,}03$, $t_+ = 0{,}05$), γιατί η μελέτη προηγήθηκε των πειραμάτων στα οποία μελετήθηκε η μέθοδος

ROUGE-W· ομολογουμένως θα ήταν καλύτερα αν η μελέτη είχε επαναληφθεί με τη μέθοδο ROUGE-W, αλλά δεν υπήρχε χρόνος για αυτό στη διάρκεια της εργασίας. Η επιλογή των κατωφλίων της μεθόδου του Γιακουμή έγινε όπως στην περίπτωση της μεθόδου ROUGE-W (βλ. ενότητα 3.2.4).

Αυτόματα Επιλεγμένες Ιδιότητες	**Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**
0	34 %
50	29 %
150	41 %
300	42 %
450	34 %
600	36 %

Τα αποτελέσματα συμφωνούν με αυτά τις προηγούμενης εργασίας [La06]. Η ορθότητα ανεβαίνει μέχρι να φτάσει σε μέγιστο στις 300 ιδιότητες και έπειτα φθίνει. Παρατηρούμε επίσης πως οι πρώτες 50 ιδιότητες δεν επηρεάζουν θετικά τα αποτελέσματα. Για περαιτέρω παρατηρήσεις σχετικές με τη φύση των ιδιοτήτων που επιλέγονται, δείτε την προηγούμενη εργασία [La06] όπου αναλύονται σε βάθος.

Ακολουθούν τα διαγράμματα που αντιστοιχούν στα παραπάνω αποτελέσματα. Σε αυτά εμφανίζονται και τα απλά συστήματα για λόγους σύγκρισης. Φαίνεται καθαρά ότι το σύστημά μας με το μέτρο του Γιακουμή (καλούμε Web-NGram αυτή τη μορφή του συστήματος) είναι καλύτερο με στατιστικά σημαντική διαφορά από τα υπόλοιπα.



Εικόνα 12: Πειράματα με μεταβλητό αριθμό αυτόματα επιλεγόμενων ιδιοτήτων

3.3.5 Συγκρίσεις συστημάτων

Εκπαιδεύσαμε το σύστημά μας χρησιμοποιώντας το μέτρο ROUGE για την αυτόματη επισημείωση παραθύρων (καλούμε Web-ROUGE αυτή τη μορφή του συστήματος) και 300 αυτόματα επιλεγμένες ιδιότητες. Τα αποτελέσματα για 50, 100, 250, 500, 1000, 1500, 2000, 2500 και 3000 ερωτήσεις εκπαίδευσης φαίνονται στον παρακάτω πίνακα.

Ερωτήσεις Εκπαίδευσης	Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)	Ορθότητα όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)	MRR
50	42,714 %	64,322 %	0,507
100	42,714 %	74,372 %	0,539
250	39,196 %	76,382 %	0,535
500	52,261 %	80,905 %	0,632
1000	47,236 %	83,417 %	0,615
1500	47,236 %	84,422 %	0,621
2000	44,221 %	81,407 %	0,583
2500	44,724 %	81,910 %	0,585

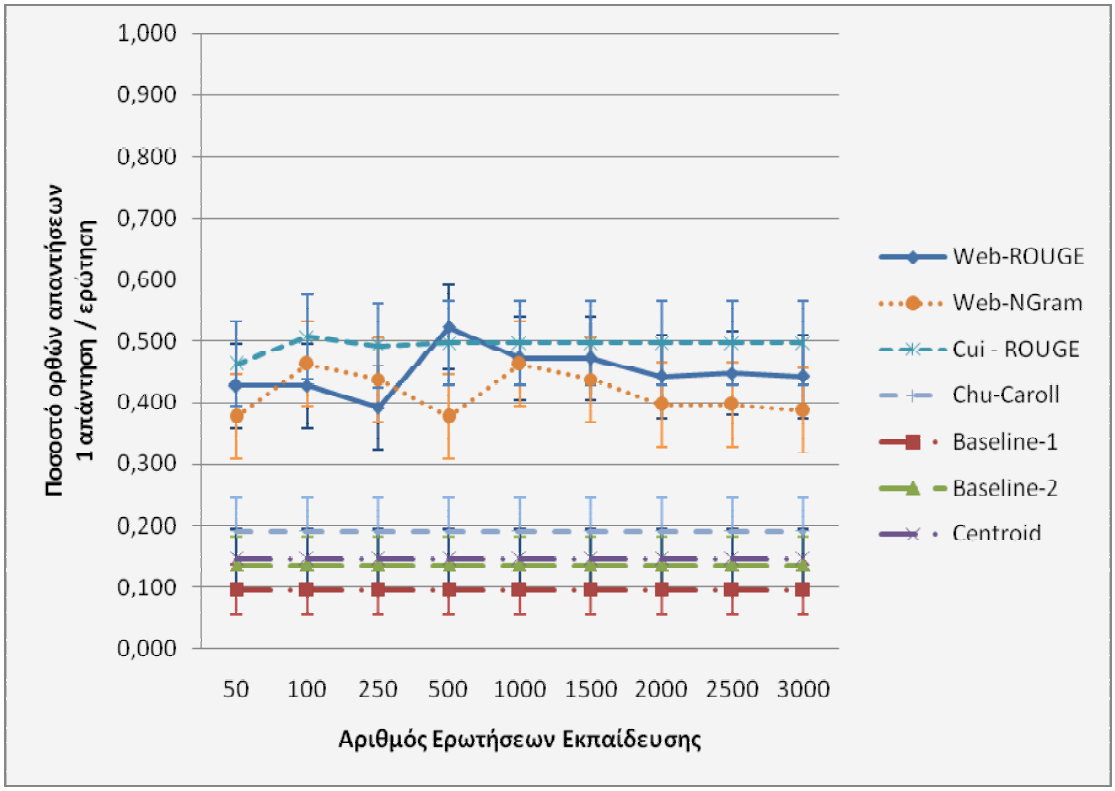
3000	44,221 %	79,397 %	0,582

Όταν επιτρέπεται μόνο μία απάντηση ανά ερώτηση, το σύστημα Web-ROUGE επιτυγχάνει την καλύτερη επίδοση στις 500 ερωτήσεις και έπειτα η επίδοσή του χειροτερεύει. Αντίθετα, όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση, η μέγιστη επίδοση επιτυγχάνεται για 1.500 ερωτήσεις εκπαίδευσης. Φαίνεται, επομένως, πως μέχρι τις 1.500 ερωτήσεις εκπαίδευσης το σύστημα κατορθώνει να εκμεταλλεύεται εν γένει τα επιπλέον δεδομένα εκπαίδευσης, αλλά μετά τις 500 ερωτήσεις δεν επιστρέφει πρώτη τη σωστή απάντηση.

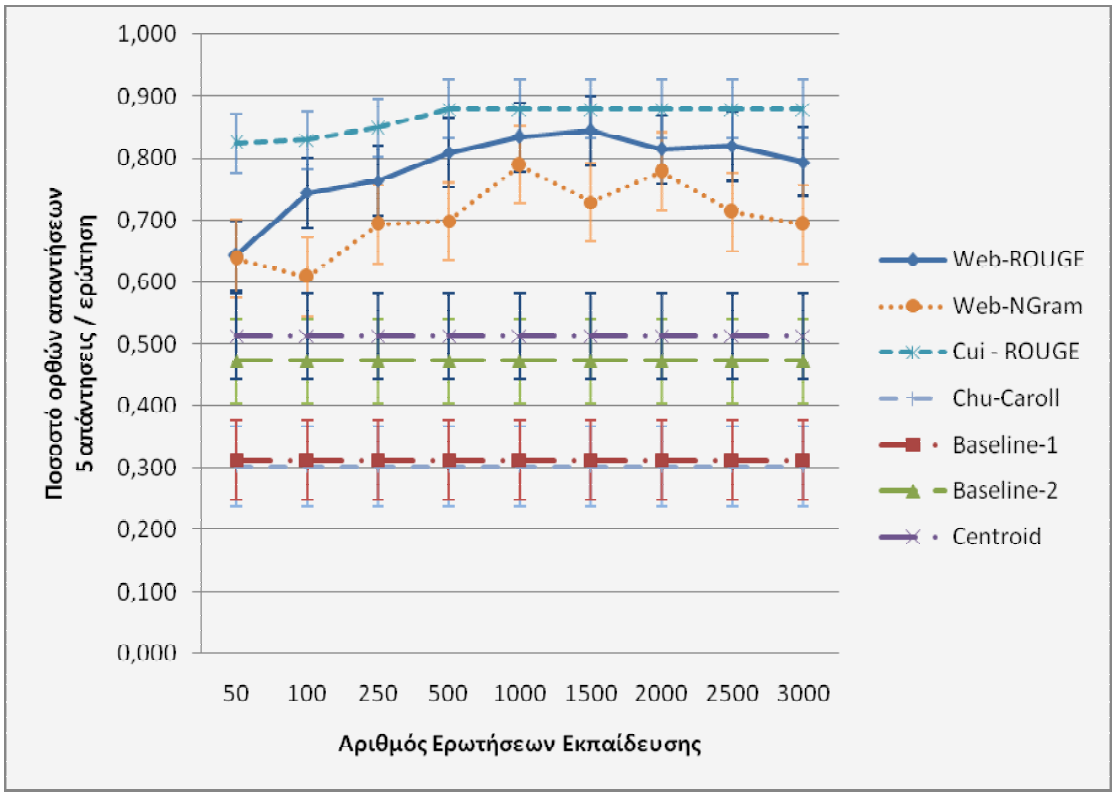
Τέλος, για λόγους σύγκρισης, εκπαιδεύσαμε και το σύστημα Web–NGram με 300 αυτόματα επιλεγμένες ιδιότητες και 50, 100, 250, 500, 1.000, 1.500, 2.000, 2.500, 3.000 ερωτήσεις εκπαίδευσης. Το σύστημα αυτό επιτυγχάνει τα καλύτερα αποτελέσματα για 1.000 ερωτήσεις εκπαίδευσης, όπως φαίνεται παρακάτω.

Ερωτήσεις Εκπαίδευσης	**Ορθότητα όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**	**Ορθότητα όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
50	37,688 %	63,819 %	0,435
100	46,231 %	60,804 %	0,413
250	43,719 %	69,347 %	0,446
500	37,688 %	69,849 %	0,499
1000	46,231 %	78,894 %	0,578
1500	43,719 %	72,864 %	0,548
2000	39,698 %	77,889 %	0,544
2500	39,698 %	71,357 %	0,494
3000	38,693 %	69,347 %	0,469

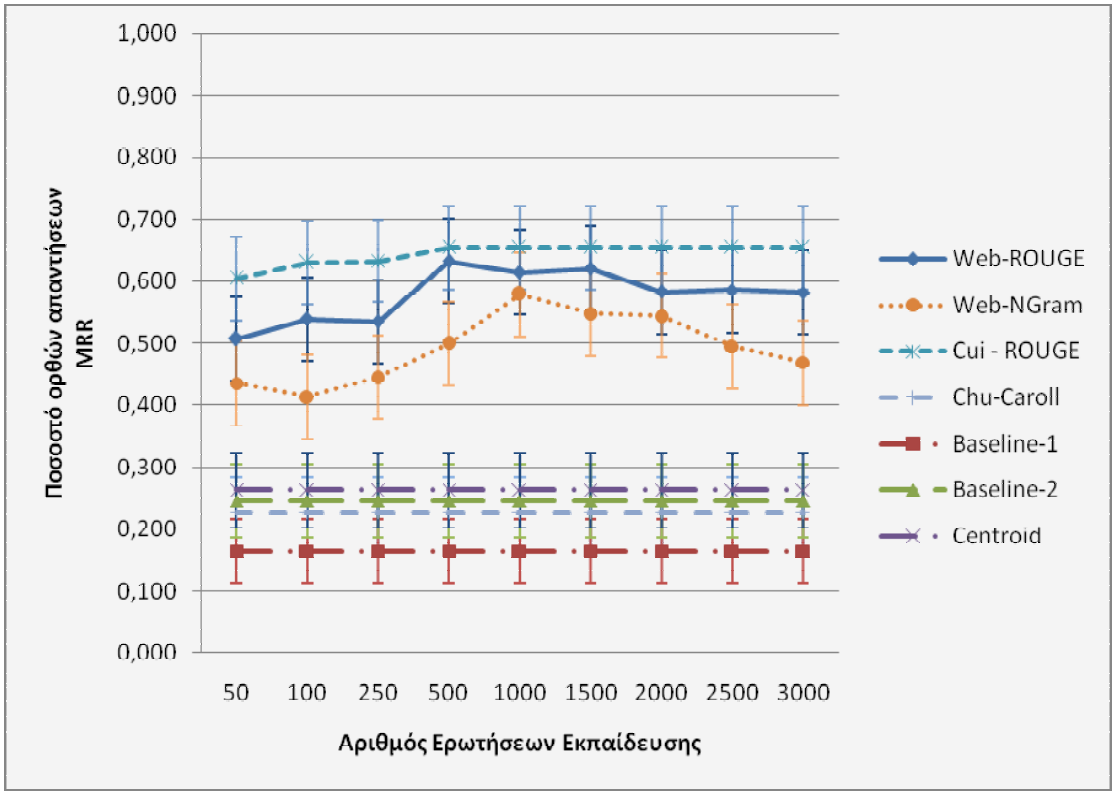
Στα παρακάτω διαγράμματα γίνεται σύγκριση ανάμεσα στα συστήματα Web–ROUGE, Web–NGram, Cui-ROUGE και των Chu-Caroll κ.ά.



Εικόνα 13: Αποτελέσματα ορθότητας όταν επιτρέπεται μία απάντηση ανά ερώτηση



Εικόνα 14: Αποτελέσματα ορθότητας όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση

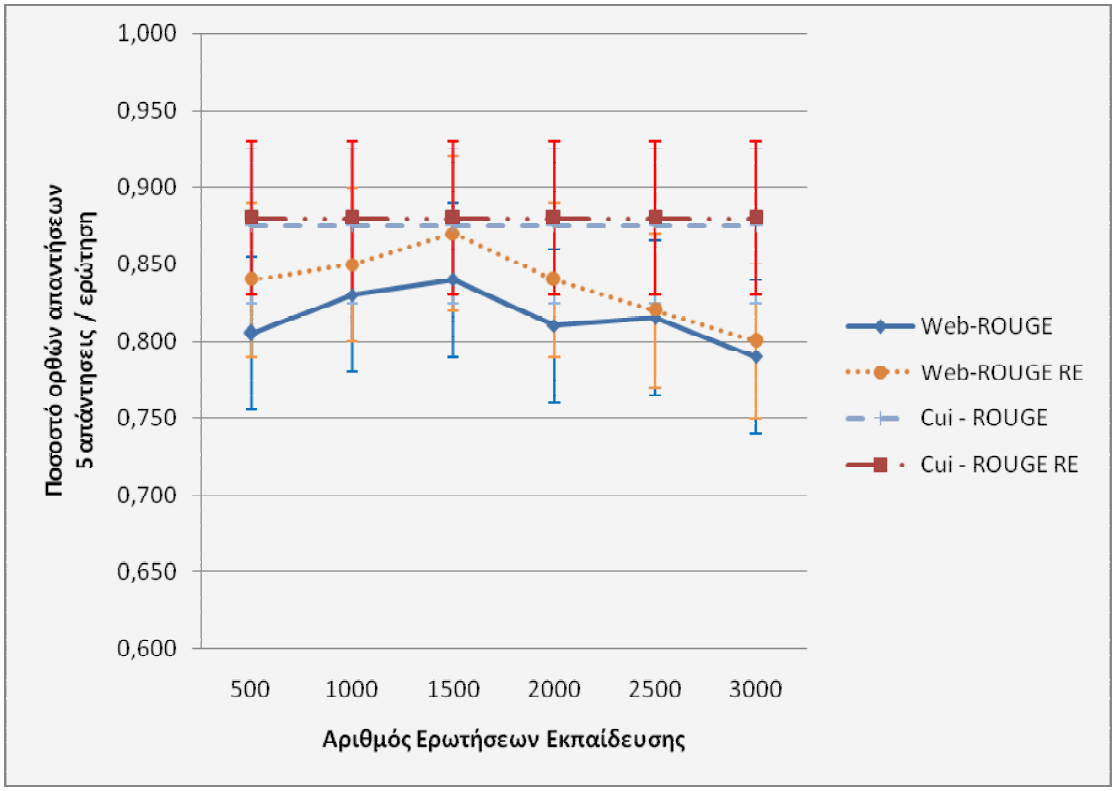


Εικόνα 15: Αποτελέσματα Mean Reciprocal Rank

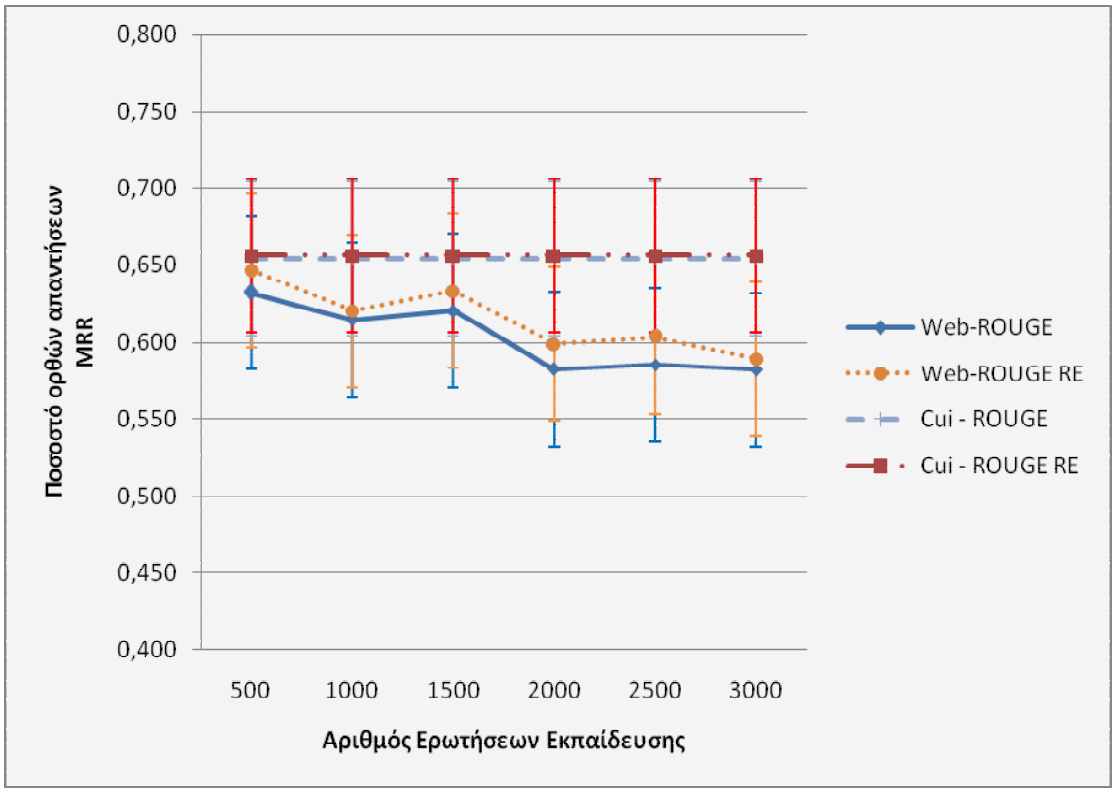
Συγκρίνοντας τα συστήματα Web-ROUGE και Web-Ngram, η πρώτη παρατήρηση είναι ότι το σύστημά μας έχει βελτιωθεί με τη χρήση του μέτρου ROUGE για την αυτόματη επισημείωση παραθύρων εκπαίδευσης, έναντι της μεθόδου του Γιακουμή, αλλά οι διαφορές είναι και πάλι μικρές . Το σύστημα των Cui κ.ά. έχει πολύ καλές και σταθερότερες επιδόσεις, αν και οι διαφορές από τα Web-ROUGE και Web-Ngram είναι σχετικά μικρές. Το συστατικό του συστήματος PIQUANT των Chu-Caroll κ.ά. που δοκιμάσαμε, αντιθέτως, παρουσιάζει πολύ χαμηλές επιδόσεις. Στην περίπτωση, μάλιστα, που επιστρέφονται πέντε απαντήσεις ανά ερώτηση, η απόδοσή του είναι χαμηλότερη εκείνης των απλών συστημάτων.

3.3.6 Πειράματα αφαίρεσης πλεονασμού

Στην παράγραφο 2.3.2 αναφέραμε δύο τεχνικές αφαίρεσης πλεονασμού από τις απαντήσεις που επιστρέφει το σύστημά μας. Συνοπτικά, η μία (Cosine) κάνει χρήση του απλού μέτρου απόστασης συνημίτονου και η δεύτερη (TF-IDF) του σταθμισμένου μέτρου απόστασης συνημίτονου με βάρη TF-IDF. Δοκιμάσαμε και τις δύο στα συστήματα Web-ROUGE και Cui-ROUGE, αλλά τα αποτελέσματά τους ήταν ακριβώς τα ίδια, οπότε στα διαγράμματα δείχνουμε μόνο μία καμπύλη (RE – redundancy elimination) και για τις δύο τεχνικές αφαίρεσης πλεονασμού.



Εικόνα 16: Αποτελέσματα ορθότητας πειραμάτων αφαίρεσης πλεονασμού



Εικόνα 17: Αποτελέσματα MRR πειραμάτων αφαίρεσης πλεονασμού

Η αφαίρεση πλεονασμού αυξάνει το ποσοστό ορθότητας του Web-ROUGE σχεδόν σε κάθε σημείο κατά 2,5%· η αύξηση στο MRR είναι 1,25. Στο σύστημα Cui-ROUGE η βελτίωση είναι πολύ μικρότερη, κατά 0,5% και 0,225 αντίστοιχα. Τα αποτελέσματα των πειραμάτων αφαίρεσης

πλεονασμού φαίνονται αναλυτικότερα στους παρακάτω πίνακες.

Ερωτήσεις Εκπαίδευσης	Web-ROUGE		Cui-ROUGE	
	Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)	**MRR**	**Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
500	84,0 %	0,647	88,0 %	0,656
1000	85,0 %	0,621	88,0 %	0,656
1500	87,0 %	0,633	88,0 %	0,656
2000	84,0 %	0,599	88,0 %	0,656
2500	82,0 %	0,603	88,0 %	0,656
3000	80,0 %	0,589	88,0 %	0,656

4. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΤΑΣΕΙΣ

Ο πρώτος βασικός στόχος της εργασίας ήταν η βελτίωση του συστήματος εντοπισμού απαντήσεων σε ερωτήσεις ορισμού που είχε κατασκευαστεί στις εργασίες των Μηλιαράκη, Γαλάνη και Γιακουμή. Ξεκινήσαμε επανεξετάζοντας τη μέθοδο αυτόματης επισημείωσης παραδειγμάτων εκπαίδευσης και προτείναμε νέες προσεγγίσεις, όπως η χρήση των μέτρων ομοιότητας του πακέτου ROUGE και η χρήση μιας δεύτερης βοηθητικής Μ.Δ.Υ. Καταλήξαμε πειραματικά στο συμπέρασμα πως η αυτόματη επισημείωση των παραδειγμάτων εκπαίδευσης με το μέτρο ομοιότητας ROUGE-W είναι η καλύτερη μέθοδος.

Ο δεύτερος βασικός στόχος της εργασίας ήταν η σύγκριση με άλλα κορυφαία συστήματα. Δυστυχώς, καταφέραμε να συγκριθούμε μόνο με το σύστημα των Cui κ.ά. και ένα συστατικό του συστήματος PIQUANT των Chu-Carroll κ.ά. Για τα υπόλοιπα συστήματα δεν υπήρχε ελεύθερα διαθέσιμη υλοποίηση, οι περιγραφές των μεθόδων τους δεν ήταν πλήρεις, ώστε να μπορέσουμε να τις υλοποιήσουμε οι ίδιοι, ή/και τα συστήματα χρησιμοποιούσαν ηλεκτρονικές εγκυκλοπαίδειες, ενώ ο σκοπός μας είναι να εντοπίζουμε ορισμούς όρων που δεν περιλαμβάνονται σε εγκυκλοπαίδειες. Επιτύχαμε, επίσης, βελτίωση των επιδόσεων του συστήματος των Cui κ.ά. επισημειώνοντας τα παραδείγματα εκπαίδευσης με το μέτρο ROUGE-W (ή το μέτρο του Γιακουμή) και τους αντίστοιχους ορισμούς εγκυκλοπαιδειών, αντί της μεθόδου του κεντροειδούς που χρησιμοποιούσαν οι Cui κ.ά. Το σύστημά μας είχε σαφώς καλύτερες πειραματικές επιδόσεις από το σύστημα των Chu-Carroll κ.ά. Οι επιδόσεις του, όμως, ήταν κατώτερες από εκείνες του (βελτιωμένου) συστήματος των Cui κ.ά., αν και οι διαφορές ήταν μικρές.

Τέλος, δοκιμάσαμε δύο απλές τεχνικές αφαίρεσης του πλεονασμού από τις απαντήσεις του συστήματός μας, όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση. Και με τις δύο τεχνικές, πετύχαμε βελτίωση στα αποτελέσματα του συστήματός μας και μικρότερη βελτίωση στα αποτελέσματα του συστήματος των Cui κ.ά. Το σύστημα των Cui κ.ά. παρέμεινε, όμως, καλύτερο.

Τα αποτελέσματα του συστήματος της εργασίας ενδέχεται να βελτιωθούν αν υιοθετήσουμε ένα «φίλτρο» που θα επιλέγει μόνο τις υποψήφιες απαντήσεις που μοιάζουν περισσότερο με το κεντροειδές όλων των υποψηφίων απαντήσεων της ερώτησης, όπως κάνουν οι Cui κ.ά., ώστε να προωθούνται μόνο αυτές οι υποψήφιες απαντήσεις στη Μ.Δ.Υ. Επίσης, το σύστημα της εργασίας μπορεί ενδεχομένως να βελτιωθεί κάνοντας πιο ελαστικό το ταίριασμα μεταξύ των παραθύρων του όρου-στόχου και των φράσεων που αντιστοιχούν στις αυτόματα επιλεγόμενες ιδιότητες. Περαιτέρω βελτίωση ενδέχεται να προέλθει από τη χρήση ενός συντακτικού αναλυτή, που θα παρέχει συντακτικές πληροφορίες για τα παράθυρα και τους ορισμούς των εγκυκλοπαιδειών που χρησιμοποιούνται κατά την εκπαίδευση. Θα μπορούσε, ακόμη, να διερευνηθεί η χρήση άλλων αλγορίθμων μάθησης, για παράδειγμα ADABOOST [FrSc99], αντί της Μ.Δ.Υ. που χρησιμοποιεί το σύστημα της εργασίας. Επίσης, αξίζει να διερευνηθούν τρόποι συνδυασμού της μεθόδου των Cui κ.ά. με τη μέθοδο της εργασίας, για παράδειγμα χρησιμοποιώντας το αποτέλεσμα της μεθόδου των Cui κ.ά. ως πρόσθετη ιδιότητα της Μ.Δ.Υ. του συστήματος της εργασίας. Τέλος, περαιτέρω έρευνα μπορεί να γίνει και στις τεχνικές αφαίρεσης πλεονασμού. Οι τεχνικές που εξετάσαμε εδώ ήταν πολύ απλές

και όπως έδειξαν τα πειράματα χρήζουν βελτίωσης.

ΑΝΑΦΟΡΕΣ

[Mi03] Σ. Μηλιαράκη, *Χειρισμός ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2003.

[Ga04] Δ. Γαλάνης, *Αυτόματη κατασκευή παραδειγμάτων εκπαίδευσης για το χειρισμό ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων που χρησιμοποιούν μηχανική μάθηση*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2004.

[Gi05] Ε. Γιακουμής, *Βελτιώσεις και περαιτέρω αξιολόγηση μεθόδου χειρισμού ερωτήσεων ορισμού για συστήματα ερωταποκρίσεων φυσικής γλώσσας*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[La06] Γ. Λάμπουρας, *Αναθεώρηση μεθόδου χειρισμού ερωτήσεων ορισμού για συστήματα ερωταποκρίσεων και μεγαλύτερης κλίμακας πειραματική αξιολόγησή της*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.

[Lu05] Γ. Λουκαρέλλι, *Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα*, μεταπτυχιακή διπλωματική εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[MaAn07] P. Malakasiotis and I. Androutsopoulos, «Learning textual entailment using SVMs and string similarity measures». Πρακτικά του *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 45th Annual Meeting of the Association for Computational Linguistics* (ACL 2007), Πράγα, Τσεχία, σελ. 42–47, 2007.

[AnRi95] I. Androutsopoulos, G.D. Ritchie, P. Thanisch, «Natural language interfaces to databases – an introduction», *Natural Language Engineering*, 1(1):29–81, Cambridge University Press, 1995.

[Vo00] E.M. Voorhees, «Overview of the TREC-9 question answering track». Πρακτικά του *9th Text Retrieval Conference* (TREC 2000), Gaithersburg, MD, ΗΠΑ, 2000.

[Vo01] E.M. Voorhees, «Overview of the TREC2001 question answering track». Πρακτικά του *10th Text Retrieval Conference* (TREC 2001), Gaithersburg, MD, ΗΠΑ, 2001.

[Cu06] H. Cui, *Soft matching for question answering*, διδακτορική διατριβή, National University of Singapore, 2006.

[CuKa05] H. Cui, M.Y. Kan, T.S. Chua, «Generic soft pattern models for definitional question answering». Πρακτικά του *28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval* (SIGIR 2005), Salvador, Βραζιλία, 2005.

[XuWe04] J. Xu, R. Weischedel, A. Licuanan, «Evaluation of an extraction-based approach to

answering definitional questions». Πρακτικά του *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l, University of Sheffield, Βρετανία, σελ. 418 – 424, 2004.

[BGMc03] S. Blair-Goldensohn, K. McKeown, A.H. Schlaikjer, «A hybrid approach for QA track definitional questions». Πρακτικά του *12th Text Retrieval Conference* (TREC 2003), Gaithersburg, MD, ΗΠΑ, 2003.

[BGMc04] S. Blair-Goldensohn, K.R. McKeown, A.H. Schlaikjer, «Answering definitional questions: a hybrid approach». Κεφάλαιο 4 στο βιβλίο *New Directions In Question Answering*, Mark Maybury (Επιμ.), AAAI Press, 2004.

[PrCh03] J. Prager, J. Chu-Carroll, K. Czuba, C. Welty, A. Ittycheriah, R. Mahindru, «IBM's PIQUANT in TREC2003». Πρακτικά του *12th Text Retrieval Conference* (TREC 2003), Gaithersburg, MD, ΗΠΑ, 2003.

[ChCz04] J. Chu-Carroll, K. Czuba, J. Prager, A. Ittycheriah, S. Blair-Goldensohn, «IBM's PIQUANT II in TREC 2004». Πρακτικά του *13th Text Retrieval Conference* (TREC 2004), Gaithersburg, MD, ΗΠΑ,, 2004.

[ChCz05] J. Chu-Carroll, K. Czuba, P. Duboue, J. Prager, «IBM's PIQUANT II in TREC 2005». Πρακτικά του *14th Text Retrieval Conference* (TREC 2005), Gaithersburg, MD, ΗΠΑ, 2005.

[HaSo06] K.S. Han, Y.I. Song, S.B. Kim, H.C. Rim, «A definitional question answering system based on phrase extraction using syntactic patterns». *IEICE Transactions on Information and Systems*, τόμος E89-D, αρ. 4, σελ.1601–1605, 2006.

[HaCh04] K.S. Han, H. Chung, S.B. Kim, Y.I. Song, J.Y. Lee, H.C. Rim, «Korea University question answering system at TREC 2004». Πρακτικά του *13th Text REtrieval Conference* (TREC 2004), Gettysburg, MD, ΗΠΑ, σελ.446–455, 2004.

[XuCa05] J. Xu, Y. Cao, H. Li, M. Zhao, «Ranking definitions with supervised learning methods». Πρακτικά του *14th International World Wide Web Conference*, Chiba, Ιαπωνία, σελ. 811–819, 2005.

[Li04] C.Y. Lin, «ROUGE: a package for automatic evaluation of summaries». Πρακτικά του *ACL Workshop Text Summarization Branches Out* (ACL 2004), Βαρκελώνη, Ισπανία, 2004.

[ChLi01] C.C. Chang, C.J. Lin, *LIBSVM: a library for Support Vector Machines*, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[AlCa20] J. Allan, J. Callan, F. Feng, D. Malin, «INQUERY at TREC8». Πρακτικά του *8h Text Retrieval Conference* (TREC 1999), Gaithersburg, MD, ΗΠΑ, 1999.

[Co95] W.W. Cohen, «Fast effective rule induction». Πρακτικά του *12th International Conference on Machine Learning* (ICML 1995), Amsterdam, Ολλανδία, σελ. 115–123, 1995.

[CaMa03] D. Carmel, Y. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer, «Searching XML documents via XML fragments». Πρακτικά του *26th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval*(SIGIR 2003), Τορόντο, Καναδάς, σελ.151–158, 2003.

[XuHu00] E. Xun, C. Huang, M. Zhou, «A unified statistical model for the identification of English BaseNP». Πρακτικά του *38th Annual Meeting of the Association for Computational Linguistics* (ACL 2000), Hong Kong, Κίνα, 2000.

[RoWa95] S. E. Robertson, S. Walker, M. M. HancockBeaulieu, M. Gatford, A. Payne, «Okapi at TREC-4». Πρακτικά του *4th Text REtrieval Conference* (TREC 1995), Gaithersburg, MD, ΗΠΑ, , National Institute of Standards and Technology, Special Publication 500-236, 1995.

[RaBo01] L. Ramshaw, E. Boschee, S. Bratus, S. Miller, R. Stone, R. Weischedel, A. Zamanian, «Experiments in multi-modal automatic content extraction». Πρακτικά του *Human Language Technology Conference*, San Diego, ΗΠΑ, 2001.

[LDC02] Linguistic Data Consortium, *ACE phase 2: information for LDC annotators*, 2002. http://www.ldc.upenn.edu/Projects/ACE2/

[TaJa97] P. Tapanainen, T. Jarvinen, «A non-projective dependency parser». Πρακτικά του *5th Conference on Applied Natural Language Processing*, Washington DC ΗΠΑ, σελ. 64–71, 1997.

[CrSh20] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*, Cambridge University Press, 2000.

[HeGr99] R. Herbrich, T. Graepel, and K. Obermayer, «*Support vector learning for ordinal regression*». Πρακτικά του *9th International Conference on Artificial Neural Networks* (ICANN99), Λονδίνο, Αγγλία, 1999.

[CoVa95] C. Cortes and V. P. Vapnik, *«*Support-vector networks*», Machine Learning*, 20(3):273–297, 1995.

[EuGl04] B. D. Eugenio, M. Glass, *«*The kappa statistic: a second look*»*, *Computational Linguistics*, 301(1):95–101, Μάρτιος 2004.

[Va98] V. P. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[FrSc99] Y. Freund, R. E. Schapire, *«*A short introduction to Boosting*»*, *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.